Stockholms universitet

Spatiotemporal Outbreak Detection

A Scan Statistic Based on the Zero-Inflated Poisson Distribution

Benjamin Kjellson

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2015:10
Matematisk statistik
Oktober 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Master Thesis **2015:10**
http://www.math.su.se

Spatiotemporal Outbreak Detection

A Scan Statistic Based on the Zero-Inflated Poisson Distribution

Benjamin Kjellson*

October 2015

Abstract

Public health authorities continuously monitor reported disease cases, looking for patterns that suggest the beginnings of an outbreak. Such analysis increasingly has to be automatized, not least due to the sheer volume of data that is generated across hospitals and clinics on a daily basis. Scan statistics are statistical methods for detecting disease outbreaks in geographic and temporal clusters, which have seen great development in the last 20 years. This thesis contributes to this development by proposing a scan statistic based on the zero-inflated Poisson (ZIP) distribution, that draws inspiration from a recent article by Cançado et al. (2014). The ZIP distribution is appropriate when some local health centers lack the facilities to diag- nose a given disease or when reported counts are biased downwards; the latter could be due to e.g. underreporting or the lack of access to medical care for uninsured individuals. The performance of the proposed ZIP scan statistic is compared to two other scan statistics, the comparison made on both simulated and real outbreak data. Results from the simulation study indicate that the proposed scan statistic outperforms the two others, being able to more accurately detect outbreaks. Furthermore, an outbreak of the diarrheal disease *cryptosporidiosis* in a German city is analyzed; this outbreak was thoroughly investigated in a recent article by Gertler et al. (2015). A final contribution of the thesis is to provide free software in the form of an R package, scanstatistics, which is available online. This package complements existing R packages for disease surveillance and outbreak detection, such as the surveillance package (Höhle et al., 2015).

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: benkjellson@gmail.com. Supervisor: Michael Höhle.

Acknowledgements

I sincerely wish to thank my supervisor Michael Höhle for his immense patience and good spirits. I also wish to thank all the other lecturers I have encountered in my studies at Stockholm University—I have learned a lot.

Contents

1 Introduction

Rapid detection of emerging disease outbreaks is of high importance to public health authorities, as an improvement of response time by weeks, days, or even hours could save the lives of many, or simply nip the outbreak in the bud. Health authorities conducting *prospective disease surveillance* hope to accomplish this feat by monitoring reported counts of disease cases or other non-diagnostic data collected at a local level, searching for spatial, temporal, or spatiotemporal *clusters* where these quantities are higher than expected. The aim of this thesis is to present and evaluate a novel method for spatial and spatiotemporal cluster detection, based on a recent article by Cançado et al. (2014). The present chapter will provide a simple motivating example of the type of problem the method tries to solve, and hence set the scene for the following chapters.

Consider the outbreak of the disease *cryptosporidiosis* that occurred in August of 2013 in the German city of Halle (Saale), following the flooding of the river Saale. This outbreak was studied in detail by Gertler et al. (2015), who identified the disease vector as *Cryptosporidium hominis*. Cryptosporidium is the microscopic parasite that causes cryptosporidiosis, a disease symptomized by watery diarrhea, stomach cramps, and vomiting (CDC, 2015). Figure 1.1 on the right shows a map of Germany and its 402 districts (*Kreise*), with the city of Halle marked in black. In each of these districts, the weekly number of cryptosporidiosis cases reported to local health authorities is relayed to the Robert Koch Institute, which performs disease surveillance at a federal level.

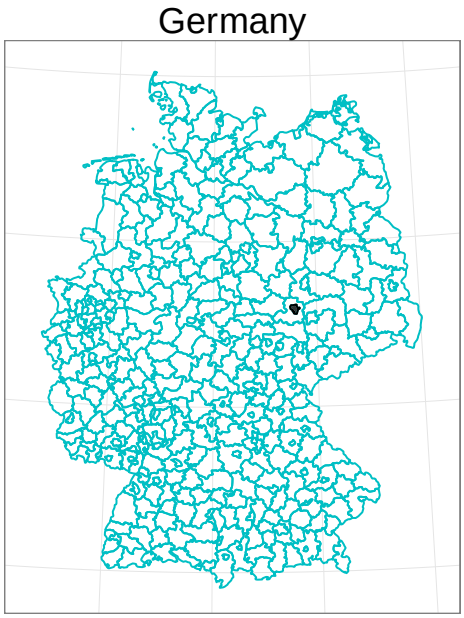

Figure 1.1 Map of the 402 districts of Germany, with the city of Halle shaded in black.

Figure 1.2 shows the time series of weekly counts of cryptosporidiosis (Cryptosporidium) reported to local health authorities in each of Germany's 402 districts, superimposed in the same plot.

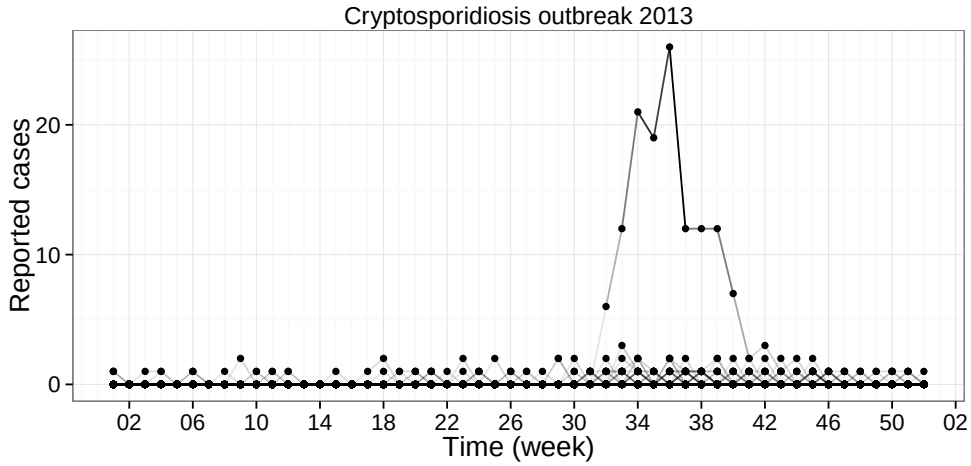

Figure 1.2 Time series of reported cryptosporidiosis cases for all 402 districts of Germany; each line is a time series for a single district, but most overlap at zero at all time points. The clear outlying line shows the reported cases for the city of Halle (Saale).

Already in week 32, it is obvious from the figure that an outbreak is emerging; as detailed by Gertler et al. (2015), the outbreak eventually lead to 167 identified cases of cryptosporidiosis in total. In conducting prospective disease surveillance, concluding that an outbreak is emerging is something we would like to do as early as possible, so that appropriate countermeasures can be taken. However, detecting an outbreak may not always be so easy. Consider for example the norovirus, also known as the winter vomiting bug: the time series of weekly counts for this disease, for all German districts, are plotted for the years 2012–2014 in Figure 1.3.

It is not at all obvious from Figure 1.3 if and when an outbreak—in the sense of unexpectedly many cases—has occurred. Cases do not seem so rare that each is noteworthy on its own, unlike the situation for cryptosporidiosis, but zero counts are still abundant. Further, there are clear seasonal patterns in the data, so that however many cases one may expect to see in each week changes over the year. When something like a flooding of a river or the contamination of locally grown and consumed produce happens, diseases may appear quickly in nearby districts. Local health authorities may not always be quick enough to see the link between cases of the same disease, particularly if the increase in cases is not suspiciously large, or if communication between different hospitals, clinics, or public health authorities is not good enough. To overcome these difficulties—to distinguish rising outbreaks from seasonal effects or expected random fluctuations, and to lessen the demands on communication between local health authorities—we wish to have statistical methods that can be employed by authorities on a national level, using the counts

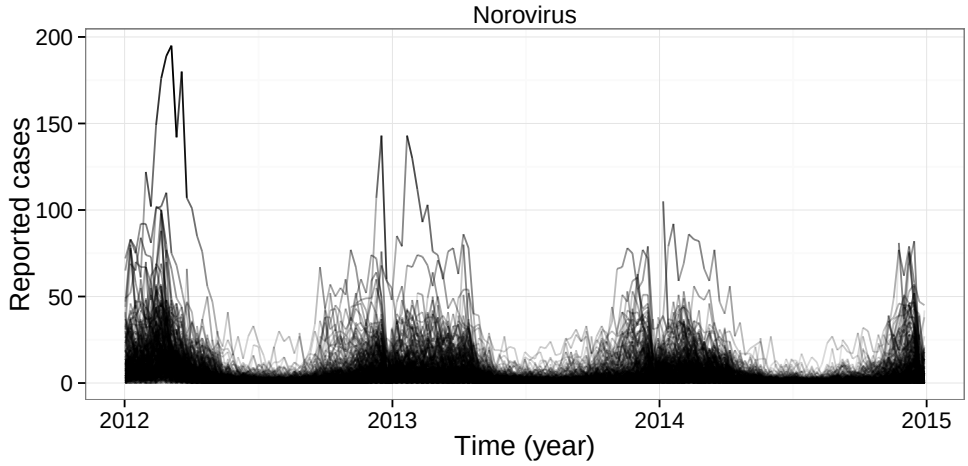

Figure 1.3 Reported cases of norovirus for all 402 districts of Germany; each line shows the weekly number of reported cases for a single district.

reported by each district. Such methods exist, and among them is a collection of methods called *scan statistics* (Glaz et al., 2009). The primary aim of this thesis is to introduce a novel scan statistic based on the recent work by Cançado et al. (2014), suitable for zero-inflated data. The use of zero-inflated distributions may be appropriate when some local health centers lack the facilities to diagnose a given disease, thus reporting counts of zero by default. These distributions may also be appropriate when reported counts are biased downwards, which could be due to e.g. underreporting or the lack of access to medical care for uninsured individuals. A secondary goal of the thesis is to provide freely accessible software that implements the proposed scan statistic and others.

1.1 Outline

The next chapter of this thesis will formalize the outbreak detection problem we wish to solve by introducing the suitable mathematical notation. Scan statistics will then be introduced as a methodology suitable for solving this problem. Chapter 2 will present an overview of scan statistics, describing in general terms the different constituents of the methodology and different variations in the actual statistical methods. Examples of two particular scan statistics will also be given with some mathematical and statistical detail. Chapter 3 then presents two other scan statistics in even further detail. The first of these is given as object of comparison to the second, which is the novel contribution of this thesis. In Chapter 4, one of the scan statistics presented in Chapter 2 and the two presented in Chapter 3 are tested on simulated data and their performances compared. In Chapter 5, a case study is conducted by applying these scan statistics to the cryptosporidiosis outbreak data mentioned previously. Lastly, a discussion of the results and merits of the proposed scan statistic is given in Chapter 6.

2 Outbreak Detection and Scan Statistics

The World Health Organization (2015) defines a disease outbreak as "the occurrence of cases of disease in excess of what would normally be expected in a defined community, geographical area or season". The outbreak detection problem considered in this thesis emphasizes the last part of that definition: if there is an outbreak, we want to know *where* it is occurring, and maximize our chances to detect it by accounting for temporal patterns in our analysis. We look for *localized* and *emerging* outbreaks, meaning those that are concentrated in a small part of a large area and that have begun recently and that are still active at present. If we can detect these types of outbreaks early, we can also stop them early, before they have spread to a wider area. As implied, the analysis is *prospective* rather than *retrospective*—our aim is not to detect outbreaks that have come and gone, but to find those that pose a threat *now* or *soon*, using the currently available data.

The section below establishes the mathematical notation that formalizes this prospective outbreak detection problem, ending with a discussion of what qualities a statistical method aimed to solve it should have. *Scan statistics* are then introduced—a collection of methods or simply a methodology whose procedures possess these qualities. The introduction will define scan statistics and provide a couple of examples which will then serve as references when the various details of the methodology are discussed, such as how hypothesis testing is conducted. With the overview of scan statistics that the introduction provides, the reader will be prepared for Chapter 3, in which the statistical methods for two particular scan statistics are presented in greater detail.

2.1 Problem Description and Notational Setup

Our starting point is a large study area, such as a country, divided into many smaller regions. These may be counties, municipalities, or other such administrative units defined by geographical boundaries. We enumerate and label these regions by $i = 1, \ldots, m$ in no particular order, so that the study area is given by $\mathcal{A} = \{1, \ldots, m\}$. Supposing that we are interested in detecting outbreaks of a specific disease, we assume that health authorities in each region i reports the number of cases y_{it} of this disease at regular (discrete) intervals of time. This implies some amount of

aggregation, as information about the exact time and geographical location of the contraction or onset of the disease is lost. As exemplified in the article by Kleinman et al. (2005) this aggregation may in some cases be mandatory due to privacy reasons (the patients' addresses were known in that study), and will most certainly simplify the computational aspects of the analysis. Furthermore, any complications due to reporting delays in the counts are ignored.

Since we look for emerging outbreaks in sequential fashion, it is convenient to count time backwards by denoting the most recent period for which we have data by $t = 1$, data from two periods ago by $t = 2$, and so on. Our interest lies in detecting outbreaks that have been active for a period of time measured in days or weeks rather than months or years. For this reason, we put an upper limit on the durations of the outbreaks, and consider only those that have a duration of say T weeks or less, if the unit of time is one week. We thus consider only outbreaks active in time intervals $I_u = [1, u]$, with $u \in \{1, \ldots, T\}$. Because we are want to find *emerging* outbreaks, which plausibly have not had the time to spread to a wider geographical area, we are justified in only trying to detect outbreaks that begin simultaneously in all (of the few) affected regions. There is another reason to restrict our search in this way: trying to detect an outbreak that starts at different time points in the different regions affected by it means a much longer computation time—time we may not have in a disease surveillance setting. In addition to the recent data we use for detecting outbreaks, we will assume that we have access to a *baseline dataset* of counts for all regions in $\mathcal{A}$. This data is assumed to be free from outbreaks, so that it can be used to estimate parameter values for the distributions of the case counts y_{it} under 'normal conditions', still accounting for regional and temporal variation in these parameters.

When a disease outbreak occurs, it manifests as higher-than-expected case counts in the regions affected by the outbreak, for its duration up to the present. The detection problem consists of concluding that there is an outbreak, and identifying the *zone* $Z \subset \mathcal{A} = \{1, \ldots, m\}$ of regions affected by it, along with its duration $u \in \{1, \ldots, T\}$. That is, we want to identify the *space-time window* (or *cluster*) $W = Z \times I_u$ containing the outbreak. Aside from the question of what qualifies as 'higher than expected', it is not at all clear how to identify the outbreak zone Z out of all 2^m possible subsets of $\mathcal{A}$. One idea would be to conduct a binary hypothesis test for each of the m regions separately—the null hypothesis stating that there is no outbreak in the region, and the alternative hypothesis stating that there is—and take as the outbreak zone those regions whose null hypotheses are rejected. However, such a procedure may well lead to a zone Z consisting of regions that are geographically scattered, which is at odds with our pursuit to find *localized* disease clusters. Further, if a classical testing procedure is used, conducting m tests and using a conventional significance level such as $\alpha = 0.05$ for each test is likely to result in a large number of false positives, as the number of regions m is often in the hundreds. Conversely, to paraphrase the argument given by Neill (2006, p. 244), counts that are two standard deviations above the mean in multiple separate regions may not seem indicative of an outbreak when considered on their own, but if these

regions all lie closely together we may well suspect that an outbreak is emerging in that area. The point is that we might gain increased detection power if we consider multiple regions jointly in our analysis, but this would seemingly aggravate the multiple testing problem as we now have not m but up to as many as 2^m zones to test. Of course, many of these 2^m zones consist of geographically dispersed regions and are thus of lesser interest, but even if we only consider zones that satisfy some proximity constraints their number may still be large enough to make an analysis computationally infeasible. What we would like is a outbreak detection methodology that 1) avoids the multiple testing problem, 2) can use the spatial information in our data, and 3) is computationally efficient. The next section gives an introduction to the scan statistics methodology, which has these qualities.

2.2 Scan Statistics

Scan statistics—a name owed to the fact that a window W is used to *scan* over the domain of interest in search of anomalous clusters—originated with Naus (1963, 1965*a*,*b*), who studied the clustering of points on a line and in the plane. Much later, the papers by Kulldorff & Nagarwalla (1995) and Kulldorff (1997) brought scan statistics closer to their current form, by formalizing the test statistics used and and generalizing the permissible shapes of the scanning windows. Scan statistics have been used not only in disease surveillance settings (as in e.g. Kulldorff, 2001; Neill et al., 2005; Takahashi et al., 2008), but has also found applications in criminology (Duczmal & Assuncao, 2004), astronomy (Moni Bidin et al., 2010), and medical imaging (Surti & Karp, 2010), to name a few examples.

Neill & Moore (2006, p. 245) give a brief description of how scan statistics are used to detect spatial or spatiotemporal clusters, which essentially amounts to the following: for each spatial or spatiotemporal window W—however these are found or selected—we assign a 'score' λ_W such that the plausibility of an outbreak in W increases with the value of the score. If we were to look at each individual window W and deem it to be an outbreak cluster if its score exceeds some fixed threshold, we would see a lot of false alarms if this threshold is set too low. On the contrary, if we set it too high our ability to detect true outbreaks diminishes, since only the truly large outbreaks (with large scores) will exceed the threshold. But since we are mainly focused on detecting a single outbreak cluster, another method becomes apparent. We define the *scan statistic* as the maximum of all scores evaluated over all possible windows, and then use the distribution of this maximum under the null hypothesis of no outbreaks in the study area to compute the probability of obtaining a value at least as large as the observed maximum. If this p-value is smaller than the chosen significance level, we identify the most likely cluster (MLC) as the window W that has the highest score. Secondary outbreak clusters, of particular interest those whose spatial components do not overlap with those of the MLC, can be found in a similar manner. In what follows, we will review—using statistical terms—the different constituents of the scan statistics methodology, supplementing the brief

explanation above by two concrete examples of scan statistics. The next sections provide the details needed to understand the statistical methods of the next chapter, and will also serve as a small survey of the literature on scan statistics.

2.2.1 Population- vs. Expectation-Based Scan Statistics

Part of the definition of an outbreak given earlier was that the counts we observe are higher than expected. But how do we determine what to expect? Neill (2006, pp. 33–35) distinguishes between two major approaches: population-based scan statistics and expectation-based scan statistics. We begin by describing and exemplifying the first kind in a purely spatial setting, as this is the setting in which population-based scan statistics are most often used, and should moreover make the spatiotemporal setting discussed next easier to understand. We then do likewise for expectation-based scan statistics, but in a spatiotemporal setting, and we put some extra detail into the example of such a scan statistic as it will serve as a base of comparison in later chapters of the thesis.

2.2.1.1 Population-Based Scan Statistics

Population-based scan statistics were proposed in the seminal article by Kulldorff & Nagarwalla (1995) and have been used in many articles since, primarily for cluster detection in the purely spatial setting (and presented as such below). Here, it is assumed that the size of the population at risk—the *denominator*—is known for each region. After a possible adjustment of this denominator for covariates such as season and whether a region is rural or urban, the null hypothesis in the population-based approach essentially states that counts in each region are generated from a distribution with mean proportional to the denominator, and that this proportionality factor is the same for all regions. To exemplify, let us consider the Poisson scan statistic devised by Kulldorff (1997), which can be considered an archetype of all scan statistics proposed in the years since. Its core assumption is that the counts $\{Y_i\}_{i=1}^m$ (the random variables corresponding to the observed counts $\{y_i\}_{i=1}^m$) are independently distributed according to a Poisson distribution; the null hypothesis is then

$$H_0 \colon Y_i \sim \text{Poisson}(q \cdot e_i), \text{ for all } i = 1, \ldots, m, \tag{2.2.1}$$

where q is the proportionality factor common for all regions, and e_i is the *known* population at risk for region i, possibly adjusted for covariates. Here and below, the mean parametrization of the Poisson distribution is used, so that if $Y \sim \text{Poisson}(\alpha)$, then $\text{P}(Y = y) = e^{-\alpha}\alpha^y/y!$, for $y = 0, 1, \ldots$.

The alternative hypothesis states that there exists a spatial zone Z for which the shared proportionality factor for the regions in it is larger than that for the regions

outside it. For *Kulldorff's scan statistic*, as it is often called, this translates to

$$H_1 : Y_i \sim \begin{cases} \text{Poisson}(q_Z e_i), & i \in Z \\ \text{Poisson}(q_{\overline{Z}} e_i), & i \in \overline{Z} \end{cases} \tag{2.2.2}$$

for some zone $Z \subset \mathcal{A} = \{1, \ldots, m\}$, with $q_Z > q_{\overline{Z}}$ and $\overline{Z}$ being the complement of Z in $\mathcal{A}$. The zone Z is here seen as an unknown parameter, and can be estimated using a profile likelihood approach: for a given Z, maximum likelihood estimates of q_Z and $q_{\overline{Z}}$ are calculated, and the likelihood is then maximized over all different zones Z considered (Patil & Taillie, 2004, pp. 185–186). To continue the example of Kulldorff's scan statistic, define the quantities $C = \sum_{i=1}^{m} y_i$ and $B = \sum_{i=1}^{m} e_i$, and let C_Z and B_Z be the corresponding sums over the regions in a given zone Z, and finally let $C_{\overline{Z}} = C - C_Z$, $B_{\overline{Z}} = B - B_Z$. Kulldorff (1997, pp. 1486–1487) then shows that the maximum likelihood estimate of q is C/B, and the MLEs of q_Z and $q_{\overline{Z}}$ are given by C_Z/B_Z and $C_{\overline{Z}}/B_{\overline{Z}}$ respectively, provided $C_Z/B_Z > C_{\overline{Z}}/B_{\overline{Z}}$. If this inequality holds, the ratio of alternative to null likelihoods for Kulldorff's scan statistic, conditional on the zone Z, is given by

$$\lambda_Z = \left(\frac{C_Z}{B_Z}\right)^{C_Z} \left(\frac{C_{\overline{Z}}}{B_{\overline{Z}}}\right)^{C_{\overline{Z}}} \left(\frac{C}{B}\right)^{-C},$$

and if the inequality does not hold, $\lambda_Z = 1$. Finally, Kulldorff's scan statistic λ^* and the most likely cluster Z^* are given by

$$\lambda^* = \max_{Z \in \mathcal{Z}} \lambda_Z, \tag{2.2.3}$$

$$Z^* = \arg\max_{Z \in \mathcal{Z}} \lambda_Z, \tag{2.2.4}$$

with $\mathcal{Z}$ being the set of all potential spatial outbreak clusters considered. How this set is chosen will be covered in section 2.2.6 below, but it is worth noting that this set could potentially consist of all 2^m subsets of the study area $\mathcal{A} = \{1, \ldots, m\}$.

The advantage of the population-based approach is that it does not demand much in the way of historical data, as we only compare the counts inside a potential cluster Z to those outside it, and the denominators are assumed to be readily obtainable. On the other hand, the detection power of this approach diminishes with the number of regions affected by the outbreak, finally unable to detect those outbreaks that affect the entire study area in a uniform manner. Neill (2006, p. 35) gives the example of a large outbreak that affects half of the study area, increasing counts there by 20%. Supposing counts remain at normal levels in the rest of the study area, the null hypothesis of no outbreak would state that counts have increased by 10% overall, so the outbreak appears smaller—and perhaps not statistically significant—in comparison to the null hypothesis (20% to 10%) than it actually is (20% to 0%). Further, the population-based approach is also susceptible to holiday effects and other situations in which counts are lower than expected in parts of the study area (according to the denominator), leading to false alarms.

2.2.1.2 Expectation-Based Scan Statistics

In some applications, the size of the population at risk is unavailable or may be inapplicable, such as when nonprescription medicine sales are used instead of case counts. In this situation, past observations could be used to forecast what values we should expect to see in each region under normal (non-outbreak) conditions, now and in the near future. In the *expectation-based* approach to scan statistics, we use historical data to estimate the mean and other parameters for each region and time point under the null hypothesis of no outbreak. We then compare the values we observe to the expected ones, and test if the counts are significantly higher than they are likely to be if there is no outbreak. Because of the clear use of time series in the expectation-based approach, this section will focus on *spatiotemporal* outbreak detection, which is the topic of this thesis. Let us exemplify again, using the *expectation-based Poisson scan statistic* proposed by Neill et al. (2005). Due to its simplicity, this scan statistic will be used as a base of comparison to the two more advanced expectation-based scan statistics presented in Chapter 3. The null hypothesis here holds that counts are independently distributed as

$$H_0 : Y_{it} \sim \text{Poisson}(\mu_{it}), \tag{2.2.5}$$

for all regions $i = 1, \ldots, m$, and all times $t = 1, \ldots, T$, with T being the maximum outbreak duration considered. Specifying the alternative hypothesis in the expectation-based approach is a bit more tricky. As the comparison for each window W is made against its past values rather than present values in the regions outside it, the null and alternative hypotheses that we test are different from those in the population-based approach. The route taken by e.g. Neill et al. (2005), Neill (2006, 2009*b*) and Tango et al. (2011) is to consider a *set* of *multiple* alternative hypotheses, each corresponding to a single space-time window $W \in \mathcal{W}$ ($\mathcal{W}$ being the set of all potential space-time outbreak clusters) and stating that an outbreak is ongoing in W. In either case, an outbreak in a window W manifests as a multiplicative increase in the mean of the counts inside W. In our example, this means that we consider the alternative hypotheses

$$H_1 : Y_{it} \sim \begin{cases} \text{Poisson}(q_W \mu_{it}), & (i,t) \in W \\ \text{Poisson}(\mu_{it}), & (i,t) \in \overline{W}, \end{cases} \tag{2.2.6}$$

for all $W \in \mathcal{W}$, with $q_W > 1$, and each count independent of others. Under the alternative hypothesis corresponding to a specific space-time window $W = Z \times I_u$, the likelihood function for a sample $\{y_{it}\}_{i=1,\ldots,m;t=1,\ldots,T}$, with corresponding mean parameters $\{\mu_{it}\}$ (the indices dropped for brevity here and henceforth) assumed known, is given by

$$L(q_W | \{y_{it}\}) = \prod_{(i,t)\in W} \exp(-q_W \mu_{it}) \frac{(q_W \mu_{it})^{y_{it}}}{y_{it}!} \times \prod_{(i,t)\in \overline{W}} \exp(-\mu_{it}) \frac{\mu_{it}^{y_{it}}}{y_{it}!}.$$

Thus, there is only one free parameter (q_W) under a given alternative hypothesis; the uncertainty in the estimate of the parameter μ_{it} is ignored. Under the null hypothesis there are no free parameters: the likelihood function is simply $L(1|\{y_{it}\})$, i.e. the same likelihood as above with $q_W = 1$. So, for this space-time window W, the distribution of the counts outside W under the alternative hypothesis will agree with that under the null hypothesis. Accordingly, the ratio of alternative to null likelihoods will cancel the likelihood contributions (factors in the likelihood) of these counts, so that the likelihood ratio conditional on W is given by

$$\lambda_W = \frac{L(q_W|\{y_{it}\})}{L(1|\{y_{it}\})} = \prod_{(i,t)\in W} \frac{\exp(-q_W\mu_{it})}{\exp(-\mu_{it})} q_W^{y_{it}}.$$

Neill (2006, pp. 36–37) derives the MLE of q_W using the likelihood function for the entire sample, under the alternative hypothesis corresponding to a given window W. An alternative argument to obtain the MLE is as follows: First, counts from outside the space-time region W are independent of those inside it and do not depend on q_W in their distributions, so are irrelevant for inference about this parameter. Second, since the counts Y_{it} for $(i,t) \in W$ are independently distributed as $Y_{it} \sim \text{Poisson}(q_W\mu_{it})$, the sum $\sum_{(i,t)\in W} Y_{it}$ is a sufficient statistic for q_W and has a Poisson distribution with mean $\sum_{(i,t)\in W} q_W\mu_{it} = q_W \sum_{(i,t)\in W} \mu_{it}$. The maximum likelihood estimator of the Poisson (mean) parameter is the sample mean; $\sum_{(i,t)\in W} y_{it}$ is thus used to estimate $q_W \sum_{(i,t)\in W} \mu_{it}$. Since $\sum_{(i,t)\in W} \mu_{it}$ is known, we see that the MLE of q_W under the restriction $q_W > 1$ must be

$$\hat{q}_W = \max\left\{1, \frac{\sum_{(i,t)\in W} y_{it}}{\sum_{(i,t)\in W} \mu_{it}}\right\}. \tag{2.2.7}$$

The test statistic for the region W is then given by the likelihood ratio

$$\lambda_W = \frac{L(\hat{q}_W|W, \{y_{it}\}, \{\mu_{it}\})}{L(1|W, \{y_{it}\}, \{\mu_{it}\})} = \prod_{(i,t)\in W} \frac{\exp(-\hat{q}_W\mu_{it})}{\exp(-\mu_{it})} \hat{q}_W^{y_{it}}.$$

This statistic is calculated for all regions $W \in \mathcal{W}$, and the scan statistic λ_W and most likely space-time cluster W^* are given by

$$\lambda^* = \max_{W\in\mathcal{W}} \lambda_W, \tag{2.2.8}$$

$$W^* = \arg\max_{W\in\mathcal{W}} \lambda_W. \tag{2.2.9}$$

Neill (2006, p. 34) notes that the expectation-based approach has a higher detection power than the population-based approach in cases where sufficient amounts of historical data exists. In particular, it is better at detecting outbreaks that affect a large part of the study area, and can account for the impact of holidays and similar events, provided these are present in the historical data.

2.2.2 Response Distributions

In the previous section, we gave two examples of scan statistics that assumed a Poisson distribution for the counts $\{y_{it}\}$. Since its appearance in the 1997 paper by Kulldorff, this distribution—attractive due to its simplicity—has found repeated use in various scan statistic formulations, for example employed in the articles by Duczmal & Assuncao (2004), Patil & Taillie (2004), Neill et al. (2005), and Tango & Takahashi (2005) just to name a few. This section will provide a brief overview of what other distributional assumptions are available in the literature on scan statistics.

Kulldorff (1997), in addition to the population-based Poisson scan statistic, also formulated a population-based scan statistic based on Bernoulli-distributed counts. Variations of this scan statistic can be found in the papers by Patil & Taillie (2004) and Christiansen et al. (2006). More recently, scan statistics based on counts that have a negative binomial distribution (Tango et al., 2011) or a zero-inflated Poisson distribution (Cançado et al., 2014) have been formulated. Scan statistics can also be defined for continuous-valued values: Neill (2006) defines two scan statistics for the normal distribution distribution with applications in neuroimaging, and Patil & Taillie (2004) does likewise for the gamma and log-normal distributions, with potential applications in environmental statistics.

The above distributions are used in a frequentist framework, in which the observed value of the scan statistic is used to obtain a p-value. The latter distribution is not available in closed form except in the most simple of cases. For this reason, most of the previously mentioned papers on scan statistics rely on Monte Carlo simulations to obtain the p-values—a topic to be covered in Section 2.2.5. These simulations can at times be computationally prohibitive. To deal with such computational issues, Neill et al. (2006) propose a 'Bayesian spatial scan statistic', which is shown to have higher power and faster runtime than frequentist alternatives such as Kulldorff's (1997) scan statistic. Similar to Kulldorff's statistic, the counts are assumed to be Poisson-distributed with an expected value that is proportional to the population at risk. The difference in this Bayesian setting is that the proportionality factor is itself assumed to be random, being gamma-distributed with different parameters depending on whether an outbreak has occurred or not. In this Bayesian approach there is no need for the Monte Carlo simulations to get a p-value. Instead, expert knowledge or historical outbreak data can be used to specify how an outbreak ought to manifest in the monitored counts, and Bayes formula can then be employed to obtain posterior probabilities for the occurrence of an outbreak in each potential cluster. The Bayesian spatial scan statistic is extended to a multivariate setting by Neill & Cooper (2010). After this overview of alternative approaches to scan statistics, the thesis will focus solely on the frequentist methodology for count data hereafter.

2.2.3 Outbreak Types

Related to the choice of response distribution of the counts is the form in which outbreaks are believed to manifest in them. In the above examples of both population- and expectation-based scan statistics, outbreaks were assumed to manifest as a multiplicative increase in the mean of the (Poisson) distribution, this increase being the same for all regions and time points affected by the outbreak. Such an outbreak cluster is referred to as a 'hot-spot' cluster by Kulldorff et al. (2003), and its simplicity has both advantages and disadvantages. Though it makes computations relatively simple, it seemingly lacks some realism in the sense that outbreaks should intuitively spread over time, and also vary in intensity over both the regions it affects and the course of its duration. Capturing such phenomena is difficult, but a few attempts have been made. Neill et al. (2005) formulates an expectation-based scan statistic for which outbreaks are assumed to have an increasing effect on the Poisson mean over the duration of the outbreak, this increase being the same for all affected spatial regions. In a more recent paper, Tango et al. (2011) derive an efficient score scan statistic for counts that are assumed to have either a Poisson or a negative binomial distribution. Here, the mean of the counts is allowed to increase according to a monotonically increasing function of the duration of the outbreak (with some restrictions). In this thesis however, we focus on the simpler hot-spot outbreak types.

2.2.4 Covariates and Parameter Estimation

When working with disease data collected at different locations and at different times of the year, we may have reason to adjust our expectations according to these factors. For example, in the summer people change dietary habits—they barbecue, letting food sit in the sun, or otherwise expose food to warmth. Bacteria such as Salmonella like these conditions, and it is therefore natural to expect a higher number of people to fall ill due to Salmonella in the summer than in the winter. For other diseases, we might also foresee some differences in the number of disease cases between regions, based on e.g. population density and whether these regions are rural or urban.

To account for regional and temporal effects in the detection of disease clusters, Kleinman et al. (2004) propose a method based on generalized linear mixed models (GLMMs). Here, a binomial logistic regression with fixed temporal effects and a random effect for each region is used to estimate the probability of a disease case at each location and time point, with the population at risk for each location and time known. In a later paper, Kleinman et al. (2005) combines the use of GLMMs with a scan statistic approach to cluster detection in order to adjust the *denominators* (populations at risk) in the study for spatial and temporal covariates. More relevant to the expectation-based Poisson scan statistic we considered earlier, Kleinman (2005) evaluates and compares the performance of both fixed effects and

mixed effects Poisson regression models. In the fixed effects model, the expected value μ_{it} of the random count Y_{it} (corresponding to the observed count y_{it}) under the null hypothesis of no outbreak is modeled as

$$\log \mu_{it} = \log \mathrm{E}[Y_{it}] = \alpha_i + \sum_j \beta_j x_{jt}, \tag{2.2.10}$$

where we see that each region has a separate (fixed) intercept, but the counts in all regions share the same coefficients for the covariates x_{jt} (which do not necessarily vary with time).

In the mixed effects model, the region-specific intercept is separated into two components: a fixed effect γ that is shared by all regions, and a random effect $a_i \sim \mathcal{N}(0, \sigma^2)$ that is specific to region i and represents the temporal variability of the counts in it. In this case, the parameter μ_{it} is the conditional expected value of y_{it}, such that

$$\log \mu_{it} = \mathrm{E}[Y_{it}|a_i] = \gamma + a_i + \sum_j \beta_j x_{jt}. \tag{2.2.11}$$

The estimators of the a_i's are referred to as 'shrinkage estimators', the estimate of the 'total' intercept $\gamma + a_i$ for a given region i improving by inclusion of the intercept term γ common to all regions. Indeed, as Kleinman (2005) remarks, the advantage of using the mixed effects model is that estimation of the intercept terms have smaller standard errors than those in the fixed effects model, and estimation improves with the number of regions m. The fixed effects approach still produces theoretically unbiased estimates, however. Lastly, Tango et al. (2011) extends the work on GLMMs for scan statistics by developing a new space-time scan statistic for counts with a negative binomial distribution parametrized by its mean and an 'overdispersion' parameter. The authors suggest that regional random effects can be included in the linear predictor—the logarithm of the mean—along with fixed effects for region and time. However, the GLMM approach is not actually applied in the data analysis section of the paper, the authors instead using a simpler moving average method for parameter estimation.

Indeed, many papers on scan statistics indicate a preference for simpler parameter estimation methods than those based on GLMs or GLMMs. Reasons for this could be that not enough relevant historical data is available, as in Tango et al. (2011, p. 108), or that fitting GLMs or GLMMs is simply too time-consuming when analyses are run daily on thousands of concurrent time series and short detection times are critical, as alluded to in Kleinman et al. (2004) and Kleinman (2005). As an example of how the expected values of counts—e.g. the μ_{it} for our expectation-based Poisson scan statistic—can be estimated when no historical data is available, consider the 'current day' method proposed by Kulldorff et al. (2005). Here, counts are assumed to be independently distributed across space and time, and the expected value for

the count in region i at time t is estimated as

$$\hat{\mu}_{it} = \frac{\left(\sum_{j=1}^{m} y_{jt}\right)\left(\sum_{k=1}^{T} y_{ik}\right)}{\sum_{j=1}^{m}\sum_{k=1}^{T} y_{jk}}. \tag{2.2.12}$$

Other examples of these relatively simple estimation methods are the time series analysis methods employed by Neill et al. (2005), which include exponentially weighted moving averages stratified or adjusted for day of week, and Holt-Winters seasonal method used by Neill (2009*b*).

2.2.5 Hypothesis Testing

In the introduction to scan statistics given above, we stated that the distribution of the scan statistic under the null hypothesis can be used to obtain a p-value corresponding to the observed value of the statistic. If this p-value is smaller than the chosen significance level the null hypothesis is rejected. Such a rejection indicates that an outbreak is ongoing in the spatial or spatiotemporal cluster (the *most likely cluster*, or MLC) corresponding to the observed scan statistic. But as it turns out, this null distribution cannot be expressed in closed analytical form for any scan statistic of practical worth. The consequence is that the p-value cannot be obtained by a simple function evaluation or approximation. The standard workaround, suggested by Kulldorff & Nagarwalla (1995) in reference to Dwass (1957), is to use Monte Carlo simulation. Kulldorff (1999, p. 309) describes the Monte Carlo hypothesis testing procedure for a scan statistic as being composed of the following four steps:

1. Calculate the value λ^{obs} of the scan statistic using the observed data, also noting the corresponding cluster.

2. Simulate a large number R of data sets, randomly generating these values from under the null hypothesis of no outbreaks.

3. For each of these replicate data sets, calculate the value of the scan statistic.

4. Reject the null hypothesis of no outbreak at significance level α if the observed value of the scan statistic is among the top 100α percent of all scan statistic values calculated (simulated and observed).

If we want an (exact) p-value, we can use the replicate scan statistics $\{\lambda_r^{\text{sim}}\}_{r=1}^{R}$ to calculate it as

$$p = \frac{1 + \sum_{r=1}^{R} \mathbb{1}\{\lambda^{\text{sim}} \geq \lambda^{\text{obs}}\}}{1 + R} \tag{2.2.13}$$

It may also be of interest to identify other potential outbreak clusters, particularly those whose spatial component does not overlap with that of the MLC. Kulldorff (1997, p. 1492) states that this can be done by ranking the likelihood ratio values of all the other potential clusters in the original data set, and comparing these to the replicate scan statistics (which are likelihood ratios). If any of these secondary clusters have a likelihood ratio value that would have caused the null hypothesis to be rejected on its own, it is seemingly an outbreak cluster. This type of test is conservative however, as a secondary cluster from the original data is compared to the most likely clusters from the simulated data sets.

A disadvantage of using Monte Carlo simulations for hypothesis testing is the computational effort involved: not only must the scan statistic be calculated for the observed data, but new data sets must be randomly generated and the statistic calculated on these. A second disadvantage is the inability of the Monte Carlo p-value to help us distinguish between clusters with very large observed values of the scan statistics, as these p-values cannot get smaller than $1/(1 + R)$. To investigate potential ways of reducing computation times and also increase the precision of the p-values, Abrams et al. (2006) compared the p-values obtained from a large number (100,000,000) of Monte Carlo replications of a spatial scan statistic, to those obtained by fitting a number of parametric distributions (Gumbel, gamma, normal, and log-normal) on a smaller number (999) of replicate scan statistics. The conclusion from this simulation study was that the p-values based on fitting a Gumbel distribution can be preferable to the Monte Carlo p-values even when they are both generated from the same number of Monte Carlo replicates. In a later paper, the same authors (Abrams et al., 2010) further investigate the usefulness of the Gumbel distribution in obtaining p-values for spatial scan statistics that are as accurate as those obtained from a larger number of Monte Carlo simulations. The results indicate that about 10 times as many simulations are needed to obtain the same (detection) power as that of the Gumbel approximation used.

Another way to reduce computation times is to use empirical values of the scan statistic, if sufficiently many of these can be calculated from past non-outbreak conditions. The current observed value of the scan statistic can then be compared to these empirical values, instead of simulated values. Neill (2009*a*) compares the performance of a number of spatial scan statistics—among these the expectation-based Poisson scan statistic—on four different data sets with 'injected' outbreaks. This means that extra counts are added on top of real-world non-outbreak data, creating a synthetic outbreak in a few regions of the study area. In addition to much-improved computation times, the results of the study by Neill indicate that the empirical approach may even lead to faster detection times and lower false positive rates.

Finally, Kulldorff (2001, p. 69) notes that for a surveillance system, whether a detected cluster should be investigated or not should not be determined simply by comparing the observed p-value to a strict significance level. The p-value should instead be seen as an indicator of the evidence of an outbreak, and the initial effort

put into an investigation of the cluster should depend on the strength of the evidence. When the evidence is strong—indicated by a small p-value—a more detailed epidemiological study using the individual cases should be carried out.

2.2.6 Cluster Shapes

In the example above, we defined the scan statistic as a maximum over a set $\mathcal{W}$, the set of all potential outbreak clusters $W = Z \times I_u$. Because we only consider outbreaks that begin at the same time in all of the regions in Z, the temporal component I_u of a space-time window W is of less importance, and we focus instead on the set $\mathcal{Z}$ of all spatial zones. This set could be as large as 2^m, the size of the power set of all regions in the study area $\mathcal{A} = \{1, \ldots, m\}$, but will due to the reasons discussed in Section 2.1 be much smaller in practice. So how is $\mathcal{Z}$ chosen or constructed, in practice?

One can distinguish between two types of methods in the literature on scan statistics. The first type uses only the spatial information of the data to construct the set $\mathcal{Z}$ of all potential outbreak zones $Z \subset \mathcal{A} = \{1, \ldots, m\}$. As an example, Kulldorff (1997) constructs the zones Z by placing circles of expanding radii centered at (the centroid of) each region $i \in \mathcal{A}$. For a given region i, the first zone Z_{i_1} is taken to contain region i only; $Z_{i_1} = \{i\}$. Next, the radius of the circle is expanded until the centroid of another region j is covered by the circle, and new zone Z_{i_2} is chosen to include both i and j; $Z_{i_2} = \{i, j\}$. This process is continued until a maximum radius is reached; one for which no more than say a half of the population at risk in the entire study area is included in the regions of the corresponding zone. Tango & Takahashi (2005) consider two other approaches of constructing $\mathcal{Z}$. The first approach is similar to the circle-expanding method above, except that for a given center region i, it is the k nearest neighbors of i that are included in each new zone, for each k from 0 up to some maximum integer K. In the second method, subsets of these k-nearest neighbor zones are also included: those subsets that include the center region i and that are connected, in the sense that each region shares a border with another region in the subset.

The second method for choosing or constructing potential clusters Z (or W) incorporates not only the spatial information of the regions in which the counts are collected, but the value of the counts themselves. The following examples provide a short overview. Duczmal & Assuncao (2004) uses a graph-based simulated annealing strategy for detecting arbitrarily shaped clusters in the spatial-only setting; the solutions obtained being quasi-optimal, but seemingly better at detecting clusters that are not in fact circular (as in Kulldorff (1997)). However, the clusters found by this method sometimes have a highly irregular shape. In a later paper, Duczmal et al. (2007) presents an approach based on a genetic algorithm, which is shown to outdo the simulated annealing method in terms of runtime, variance, and detection power. Assuncao et al. (2006) presents two graph theory approaches based

on minimum spanning trees (MST) which are relatively fast. One of the methods include the *upper level set* windows of Patil & Taillie (2004) as a special case, but this method has low power, and the other method tends to overestimate cluster sizes when the true clusters are of some particular shapes. Duczmal et al. (2011) also proposes a MST-based approach, but with a novel distance metric.

In this thesis, we consider only the k-nearest neighbor approach. This is because the scan statistics that we consider in this thesis—the expectation based Poisson scan statistic considered in Section 2.2.1.2 above and the two scan statistics presented in the next chapter—are of the first type described above: they only rely on the spatial information of the data to construct $\mathcal{Z}$. When we compare these methods in chapters 4 and 5, the shape of the spatial component of the outbreak will be under our control, and the k-nearest neighbor approach is an attractive alternative due to its simplicity. In settings where it is unknown what spatial shape an outbreak will take, other approaches could be used.

This chapter has given an overview of the different constituents of the scan statistics methodology. In the next chapter, we present the mathematical details of two expectation-based scan statistics—one of which is a novel contribution of this thesis—that are more advanced than the one given as an example in this chapter.

3 Advanced Statistical Methods

The previous chapter introduced the concept of scan statistics in a statistical context and gave the reader a specific example in the expectation-based Poisson scan statistic. This particular scan statistic serves as a good point of reference due to its simplicity. However, more advanced methods are available. In this chapter we consider two other expectation-based scan statistics, the first being the efficient score scan statistic developed by Tango et al. (2011), which can accommodate a negative binomial distribution for the counts. The second is a novel scan statistic inspired by the population-based ZIP scan statistic proposed by Cançado et al. (2014). This method assumes that the distribution generating the case counts can be approximated by a zero-inflated Poisson distribution, and uses the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) to obtain parameter estimates. In the next couple of sections, we introduce these scan statistics and their advantages and disadvantages over the expectation-based Poisson statistic.

3.1 A Negative-Binomial Score Scan Statistic

When count data has larger variance than expected value, use of the Poisson distribution for detecting disease outbreaks may lead to a high number of false alarms. Counts sufficiently larger than the estimated mean might seem strongly indicative of an outbreak under the Poisson assumption, while they might be entirely within what is to be expected if some other more dispersed distribution was used, if appropriate. Such problems of overdispersion make it natural to consider the negative binomial distribution, which allows the variance to be larger than the mean. For the expectation-based Poisson statistic considered in the previous chapter, we posited that an outbreak affecting a space-time window W would have a multiplicative effect q_W on the expected values for the counts in this window. In Equation (2.2.7) we were then able to derive an analytical expression for the maximum likelihood estimate of this factor q_W and use it to calculate a likelihood ratio statistic in closed form for each potential outbreak cluster W. If we were to consider the same type of outbreak model for counts with a negative binomial response distribution, we would not be able to derive a closed form expression for this factor q_W. Yet this difficulty can be overcome—and q_W even allowed to take a more general form than a constant—as shown in the paper by Tango et al. (2011). Their starting point is to consider counts Y_{it} that follow a negative binomial distribution, the null hypothesis

stating that

$$H_0 \colon Y_{it} \sim \mathrm{NB}(\mu_{it}, \phi_{it}). \tag{3.1.1}$$

Here, the mean parametrization of this distribution is used, so that

$$\mathrm{P}(Y_{it} = y_{it}) = \frac{\Gamma(\phi_{it} + y_{it})}{\Gamma(\phi_{it}) y_{it}!} \left(\frac{\phi_{it}}{\phi_{it} + \mu_{it}} \right)^{\phi_{it}} \left(\frac{\mu_{it}}{\phi_{it} + \mu_{it}} \right)^{y_{it}}, \tag{3.1.2}$$

$$\mathrm{E}[Y_{it}] = \mu_{it}, \tag{3.1.3}$$

$$\mathrm{Var}(Y_{it}) = \mu_{it} + \mu_{it}^2/\phi_{it}. \tag{3.1.4}$$

The value of μ_{it} and ϕ_{it} are assumed to be known, but are in practice estimated from historical data. Estimation and regression techniques for the negative binomial distribution are developed in Lawless (1987); Hilbe (2011) provides a more up-to-date treatment, discussing various parameterizations and software implementations.

Below, we take a look at two scan statistics proposed by Tango et al. (2011) based on the negative binomial distribution and two different types of outbreaks. The contribution of this thesis with respect to these scan statistics is to provide a greater mathematical detail in their derivation, compared to what was given by Tango et al.. We begin by first considering the scan statistic for a hot-spot cluster model of an outbreak, as was described in Section 2.2.3.

3.1.1 Hot-spot Cluster Model

Just as for the expectation-based Poisson scan statistic, a separate alternative hypothesis is considered for each potential cluster W. Tango et al. (2011) consider two types of outbreak models, the simplest being the *hot-spot cluster model* introduced by name in Section 2.2.2. For a given space-time window W, the alternative hypothesis states that

$$H_1 \colon Y_{it} \sim \begin{cases} \mathrm{NB}(q_W \mu_{it}, \phi_{it}), & (i,t) \in W \\ \mathrm{NB}(\mu_{it}, \phi_{it}), & (i,t) \in \overline{W} \end{cases}, \tag{3.1.5}$$

with $q_W > 1$ and where $\overline{W}$ is the complement of W, i.e. the regions and time points outside this window. As explained in Section 2.2.1.2 this means that we have a set of alternative hypotheses, one for each $W \in \mathcal{W}$. In this section, we derive a scan statistic corresponding this outbreak model. It is worth noting that these derivations are not given in the paper by Tango et al. (2011), which only gives the final result, but rather represent an original contribution of this thesis.

The idea here is to use Rao's score test statistic, which involves evaluating the score function and the Fisher information derived from the log-likelihood function of the

above model, at the value of q_W under the null hypothesis. To this end, we first form the log-likelihood for a given space-time cluster $W = Z \times I_u$. Since q_W is the only free parameter under the alternative hypothesis (the μ_{it}'s and ϕ_{it}'s already given), we see that when replacing μ_{it} by $q_W\mu_{it}$ in Equation (3.1.2), the log-likelihood is (up to an additive constant)

$$\ell_W(q_W|\{y_{it}\}) = \sum_{(i,t)\in W} [y_{it}\log(q_W) - (\phi_{it} + y_{it})\log(\phi_{it} + q_W\mu_{it})] \tag{3.1.6}$$

By differentiation w.r.t. q_W of the log-likelihood, the score function and observed Fisher information are seen to be

$$U_W(q_W) = \ell'_W(q_W|\{y_{it}\}) = \sum_{(i,t)\in W} \left[\frac{y_{it}}{q_W} - \frac{(\phi_{it} + y_{it})\mu_{it}}{\phi_{it} + q_W\mu_{it}}\right], \tag{3.1.7}$$

$$\mathcal{J}(q_W) = -\ell''_W(q_W|\{y_{it}\}) = \sum_{(i,t)\in W} \left[\frac{y_{it}}{q_W^2} - \frac{(\phi_{it} + y_{it})\mu_{it}^2}{(\phi_{it} + q_W\mu_{it})^2}\right]. \tag{3.1.8}$$

Tango et al. (2011) define the parameter $w_{it} = 1 + \mu_{it}/\phi_{it}$, which they call the 'overdispersion parameter' of the negative binomial distribution. Using this parameter, we see that evaluating the score function at $q_W = 1$ gives

$$\begin{aligned} U_W(1) &= \sum_{(i,t)\in W} \left[y_{it} - \frac{\phi_{it}\mu_{it} + y_{it}\mu_{it}}{\phi_{it} + \mu_{it}}\right] \\ &= \sum_{(i,t)\in W} \frac{y_{it} - \mu_{it}}{\phi_{it} + \mu_{it}}\phi_{it} \\ &= \sum_{(i,t)\in W} \frac{y_{it} - \mu_{it}}{w_{it}} \end{aligned}$$

Similarly, the Fisher information evaluated at $q_W = 1$ becomes

$$\begin{aligned} \mathcal{I}_W(1) &= \mathrm{E}[\mathcal{J}_W(1)] \\ &= \sum_{(i,t)\in W} \left[\mu_{it} - \frac{(\phi_{it} + \mu_{it})\mu_{it}^2}{(\phi_{it} + \mu_{it})^2}\right] \\ &= \sum_{(i,t)\in W} \frac{\mu_{it}}{\phi_{it} + \mu_{it}}\phi_{it} \\ &= \sum_{(i,t)\in W} \frac{\mu_{it}}{w_{it}}. \end{aligned}$$

The score statistic, corresponding to that of Rao's score test (see e.g. Lehmann & Romano, 2008, p. 511), is then given by

$$\lambda_W = \frac{U_W(1)}{\sqrt{\mathcal{I}_W(1)}} \tag{3.1.9}$$

and has a standard normal distribution asymptotically. However, this asymptotic normality is not used in the paper by Tango et al. (2011). Instead, the maximum of this statistic over all space-time windows $W \in \mathcal{W}$ is taken, so that the scan statistic—and corresponding *most likely cluster*—are given by

$$\lambda^* = \max_{W \in \mathcal{W}} \lambda_W, \tag{3.1.10}$$

$$W^* = \arg\max_{W \in \mathcal{W}} \lambda_W. \tag{3.1.11}$$

One reason for not using the asymptotic normality of each of the λ_W's is that the number of counts that go into calculating this statistic can often be quite small—it is the number of regions in the spatial component of the space-time window W times the outbreak duration considered—and these counts may themselves be low. In the simulation study conducted by Tango et al. (2011), the largest windows W considered have a maximum of 10 separate regions and a maximum temporal length of two weeks (one count per week). Next we consider the second scan statistic in the paper by Tango et al. (2011), which corresponds to a more complex—and perhaps realistic—outbreak scenario.

3.1.2 Emerging Outbreak Model

The hot-spot model presented in the previous section assumes that the multiplicative increase in the mean of the count distribution is the same across all affected spatial regions of the outbreak, as well as over time. It may seem more natural—particularly in the context of infectious diseases—that the mean might increase over time during an outbreak, at least initially. To detect such emerging outbreaks, the second model proposed by Tango et al. (2011)—the *emerging outbreak* model—assumes that the factor by which the counts increase is a function instead of a constant. In this section, we present the derivation of the scan statistic for this model, with greater mathematical detail than what was given in the paper by Tango et al. (2011). The presentation also differs due to the choice of counting time backwards in this thesis—in Tango, Takahashi & Kohriyama's paper, time is counted in the forward direction.

In the emerging outbreak model, we consider the alternative hypothesis

$$H_1 : Y_{it} \sim \begin{cases} \text{NB}(q_{it}\mu_{it}, \phi_{it}), & (i,t) \in W \\ \text{NB}(\mu_{it}, \phi_{it}), & (i,t) \in \overline{W} \end{cases}, \quad \text{where} \tag{3.1.12}$$

$$q_{it} = h(\tau + \beta_W(u + 1 - t)), \tag{3.1.13}$$

for each space-time window $W = Z \times I_u$. Again, as in Section 2.2.1.2, this means that we have a set of alternative hypotheses, one for each $W \in \mathcal{W}$. The function $h : \mathbb{R} \to \mathbb{R}$ is assumed to be monotonically increasing over the duration of the outbreak, but as we measure time backwards in units of time ago from the present,

the function is monotonically decreasing in our notation. It is assumed that this function has finite first and second derivatives, and that the value of the function $h(\cdot)$ one unit of time before the start of the outbreak, i.e. at time $u+1$, is one. That is, $h(\tau) = 1$. The *initial slope* of the outbreak is given by

$$\left.\frac{\partial q_{it}}{\partial t}\right|_{t=u} = -\beta_W h'(\tau). \tag{3.1.14}$$

This reduces the alternative hypothesis to the statement that $\beta_W > 0$. Tango et al. (2011) show that even if the functional form of $h(\cdot)$ is unknown, and a maximum likelihood of β_W cannot be obtained analytically, the hypotheses can be tested by deriving a score test statistic similarly to what was done in Section 3.1.1 above. Starting with the log-likelihood, this is now given by

$$\ell_W(\beta_W|\{y_{it}\}) = \sum_{(i,t)\in W} \left[y_{it}\log(q_{it}) - (\phi_{it}+y_{it})\log(\phi_{it}+q_{it}\mu_{it})\right], \tag{3.1.15}$$

where $q_{it} = h(\tau + \beta_W(u+1-t))$. Differentiating with respect to β_W, we get the score function

$$\begin{aligned}
U_W(\beta_W) &= \sum_{(i,t)\in W}\left[\frac{y_{it}}{q_{it}}\frac{\partial q_{it}}{\partial \beta_W} - (\phi_{it}+y_{it})\frac{\mu_{it}}{\phi_{it}+q_{it}\mu_{it}}\frac{\partial q_{it}}{\partial \beta_W}\right] \\
&= \sum_{(i,t)\in W}\left[\frac{y_{it}}{q_{it}} - (\phi_{it}+y_{it})\frac{\mu_{it}}{\phi_{it}+q_{it}\mu_{it}}\right]\frac{\partial q_{it}}{\partial \beta_W} \\
&= \sum_{(i,t)\in W}\left[\frac{y_{it}}{q_{it}} - (\phi_{it}+y_{it})\frac{\mu_{it}}{\phi_{it}+q_{it}\mu_{it}}\right](u+1-t)h'(\tau+\beta_W(u+1-t)).
\end{aligned}$$

Letting $\beta_W = 0$ so that $q_{it}|_{\beta_W=0} = h(\tau) = 1$, and recognizing the similarity of the expression to the score function for the hot-spot cluster model, for which we switched to the parametrization using $w_{it} = 1 + \mu_{it}/\phi_{it}$, we get

$$\begin{aligned}
U_W(0) &= \sum_{(i,t)\in W}\left[y_{it} - (\phi_{it}+y_{it})\frac{\mu_{it}}{\phi_{it}+\mu_{it}}\right](u+1-t)h'(\tau) \\
&= h'(\tau)\sum_{(i,t)\in W}\frac{y_{it}-\mu_{it}}{w_{it}}(u+1-t).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\mathcal{J}_W(\beta_W) &= \sum_{(i,t)\in W}\left[\frac{y_{it}}{q_{it}^2}\frac{\partial q_{it}}{\partial \beta_W} - (\phi_{it}+y_{it})\frac{\mu_{it}^2}{(\phi_{it}+q_{it}\mu_{it})^2}\frac{\partial q_{it}}{\partial \beta_W}\right]\frac{\partial q_{it}}{\partial \beta_W} \\
&\quad - \sum_{(i,t)\in W}\left[\frac{y_{it}}{q_{it}} - (\phi_{it}+y_{it})\frac{\mu_{it}}{\phi_{it}+q_{it}\mu_{it}}\right]\frac{\partial^2 q_{it}}{\partial \beta_W^2} \\
&= \sum_{(i,t)\in W}\left[\frac{y_{it}}{q_{it}^2} - (\phi_{it}+y_{it})\frac{\mu_{it}^2}{(\phi_{it}+q_{it}\mu_{it})^2}\right]\left(\frac{\partial q_{it}}{\partial \beta_W}\right)^2 \\
&\quad - \sum_{(i,t)\in W}\left[\frac{y_{it}}{q_{it}} - (\phi_{it}+y_{it})\frac{\mu_{it}}{\phi_{it}+q_{it}\mu_{it}}\right]\frac{\partial^2 q_{it}}{\partial \beta_W^2} \\
&= \sum_{(i,t)\in W}\left[\frac{y_{it}}{q_{it}^2} - (\phi_{it}+y_{it})\frac{\mu_{it}^2}{(\phi_{it}+q_{it}\mu_{it})^2}\right]\left[(u+1-t)h'(\tau+\beta_W(u+1-t))\right]^2 \\
&\quad - \sum_{(i,t)\in W}\left[\frac{y_{it}}{q_{it}} - (\phi_{it}+y_{it})\frac{\mu_{it}}{\phi_{it}+q_{it}\mu_{it}}\right](u+1-t)^2 h''(\tau+\beta_W(u+1-t)).
\end{aligned}$$

The Fisher information at $\beta_W = 0$ thus becomes

$$\begin{aligned}
\mathcal{I}_W(0) &= \mathrm{E}\left[\sum_{(i,t)\in W}\left[y_{it} - (\phi_{it}+y_{it})\frac{\mu_{it}^2}{(\phi_{it}+\mu_{it})^2}\right](u+1-t)^2(h'(\tau))^2\right] \\
&\quad - \mathrm{E}\left[\sum_{(i,t)\in W}\left[y_{it} - (\phi_{it}+y_{it})\frac{\mu_{it}}{\phi_{it}+\mu_{it}}\right](u+1-t)^2 h''(\tau)\right] \\
&= \sum_{(i,t)\in W}\left[\mu_{it} - \frac{\mu_{it}^2}{\phi_{it}+\mu_{it}}\right](u+1-t)^2(h'(\tau))^2 \\
&\quad - \sum_{(i,t)\in W}\left[\mu_{it}-\mu_{it}\right](u+1-t)^2 h''(\tau) \\
&= (h'(\tau))^2 \sum_{(i,t)\in W}\left[\mu_{it} - \frac{\mu_{it}^2}{\phi_{it}+\mu_{it}}\right](u+1-t)^2 \\
&= (h'(\tau))^2 \sum_{(i,t)\in W}\frac{\mu_{it}}{w_{it}}(u+1-t)^2
\end{aligned}$$

It then becomes clear that when forming the score statistic λ_W as the ratio of the score function to the (signed) square root of the Fisher information, the factor $h'(\tau)$ will cancel, ridding the statistic of dependence on the functional form of $h(\cdot)$. The score statistic becomes

$$\lambda_W = \frac{\sum_{(i,t)\in W}(y_{it}-\mu_{it})(u+1-t)/w_{it}}{\sqrt{\sum_{(i,t)\in W}\mu_{it}(u+1-t)^2/w_{it}}}.$$

This statistic is calculated for each potential space-time outbreak cluster $W = Z \times I_u$, and the scan statistic λ^* and the corresponding most likely cluster W^* are given by

$$\lambda^* = \max_{W \in \mathcal{W}} \lambda_W, \tag{3.1.16}$$

$$W^* = \arg\max_{W \in \mathcal{W}} \lambda_W. \tag{3.1.17}$$

3.2 An Expectation-Based ZIP+EM Scan Statistic

As was mentioned earlier, overdispersion can sometimes be a problem when the Poisson distribution is used to model count data. Sometimes, this overdispersion is due to an overabundance of zero counts, which—when dealing with disease data—could be due to factors such as underreporting. One possible way to account for this particular feature is to use a zero-inflated model. In such a model, the counts can be viewed as generated from a mixture of two distributions: with probability p, an *excess zero* is generated from the distribution degenerate at zero, and with probability $1 - p$, a count is generated from some other distribution, such as the Poisson distribution. In particular, the Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) models have previously been used in infectious disease studies. For example, the ZIP model was used by Vergne et al. (2014) to model avian influenza outbreaks in Thailand during the 2005 outbreak, and the ZINB was used by Carrel et al. (2010) in the study of cholera outbreaks associated with floodings in rural Bangladesh.

In a recent paper by Cançado et al. (2014), a population-based scan statistic for ZIP-distributed counts is presented in the purely spatial setting. From an inference point of view, they adapt the EM algorithm methods developed by Lambert (1992) for the ZIP-distribution to arrive at maximum likelihood estimates of the relevant parameters. In this section of the thesis, we present an novel expectation-based version of this ZIP scan statistic, which also uses the EM algorithm to obtain parameter estimates. This scan statistic is the main, original contribution of this thesis. The ZIP scan statistic uses (or can use) a regression approach to model both the Poisson means and the structural zero probabilities, whereas Cançado et al. (2014) only made note of the possibility to do so for the excess zero probability. The scope is also extended from the spatial to the spatiotemporal setting. Below, we present the new scan statistic in a format similar to that given in the paper by Cançado et al. (2014).

3.2.1 Response Distributions and Hypotheses

For the expectation-based ZIP scan statistic, we assume that the number of cases Y_{it} in region i at time t follows a zero-inflated Poisson distribution, independent of

the number of cases at other times and places. The null hypothesis of no outbreak states that

$$H_0 \colon Y_{it} \sim \mathrm{ZIP}(p_{it}, \mu_{it}). \tag{3.2.1}$$

The ZIP distribution is specified by the probability mass function

$$\mathrm{P}(Y_{it} = y_{it}) = \begin{cases} p_{it} + (1 - p_{it})e^{-\mu_{it}}, & y_{it} = 0 \\ (1 - p_{it})e^{-\mu_{it}} \frac{\mu_{it}^{y_{it}}}{y_{it}!}, & y_{it} = 1, 2, \ldots, \end{cases} \tag{3.2.2}$$

which can be seen as a mixture of a degenerate distribution at zero and a Poisson(μ_{it}) distribution. The expected value and variance of Y_{it} are given by

$$\mathrm{E}[Y_{it}] = (1 - p_{it})\mu_{it}, \tag{3.2.3}$$

$$\mathrm{Var}(Y_{it}) = \mathrm{E}[Y_{it}] + \frac{p_{it}}{1 - p_{it}} \mathrm{E}[Y_{it}]^2, \tag{3.2.4}$$

from which it is seen that $\mathrm{Var}(Y_{it}) > \mathrm{E}[Y_{it}]$ when $p_{it} \in (0, 1)$. Thus, the zero-inflated Poisson distribution indeed has a larger variance than the (regular) Poisson distribution. For outbreaks, we again employ the hot-spot cluster model by considering, for each space-time window $W = Z \times I_u$, the alternative hypothesis

$$H_1 \colon Y_{it} \sim \begin{cases} \mathrm{ZIP}(p_{it}, q_W \mu_{it}), & (i, t) \in W \\ \mathrm{ZIP}(p_{it}, \mu_{it}), & (i, t) \in \overline{W}, \end{cases} \tag{3.2.5}$$

with $q_W > 1$. As in Section 2.2.1.2, this means that we have a set of alternative hypotheses, one for each $W \in \mathcal{W}$. To test these hypotheses, we will want to define a scan statistic as the maximum of likelihood ratios over all potential outbreak clusters W. So let us have a look at the likelihood function for the parameters of a ZIP distribution next.

3.2.2 Likelihoods

Suppose we are given a data set $\{y_{it}\}$ of counts and that we have been able to use historical data to estimate the corresponding parameters $\{p_{it}\}$ and $\{\mu_{it}\}$, which we believe are accurate in non-outbreak conditions. Hence, we ignore any estimation uncertainty for these parameters. Under the alternative hypothesis corresponding to given space-time window $W = Z \times I_u$, the log-likelihood function in this scenario

is

$$\begin{aligned}
\ell(q_W|\{y_{it}\}) &= \log\left(\left[\prod_{(i,t)\in W} \mathrm{P}(Y_{it}=y_{it})\right]\left[\prod_{(i,t)\in \overline{W}} \mathrm{P}(Y_{it}=y_{it})\right]\right) \\
&= \sum_{\substack{(i,t)\in W \\ y_{it}=0}} \log\left(p_{it}+(1-p_{it})e^{-q_W\mu_{it}}\right) \\
&\quad + \sum_{\substack{(i,t)\in W \\ y_{it}>0}} \left[\log(1-p_{it}) - q_W\mu_{it} + y_{it}\log(q_W\mu_{it}) - \log(y_{it}!)\right] \\
&\quad + \sum_{\substack{(i,t)\in \overline{W} \\ y_{it}=0}} \log\left(p_{it}+(1-p_{it})e^{-\mu_{it}}\right) \\
&\quad + \sum_{\substack{(i,t)\in \overline{W} \\ y_{it}>0}} \left[\log(1-p_{it}) - \mu_{it} + y_{it}\log(\mu_{it}) - \log(y_{it}!)\right].
\end{aligned}$$

Note that no analytical solution for the q_W that maximizes this likelihood can be found. To overcome this problem, we adopt an EM-algorithm approach similar to what was done by Cançado et al. (2014). First, assume that we *do* know whether the zeros in the data are excess zeros or not. Let δ_{it} be a indicator variable that takes the value 1 if the count in region i at time t is an excess zero (which happens with probability p_{it}), and takes the value 0 if it is not (which happens with probability $1-p_{it}$). Given the complete data, it can then be gathered that

$$\begin{aligned}
\mathrm{P}(Y_{it}=y_{it}|\delta_{it}=0) &= e^{-q_W\mu_{it}}\frac{(q_W\mu_{it})^{y_{it}}}{y_{it}!}, \quad y_{it}=0,1,\ldots, \\
\mathrm{P}(Y_{it}=0|\delta_{it}=1) &= 1, \\
\mathrm{P}(Y_{it}>0|\delta_{it}=1) &= 0,
\end{aligned}$$

for $(i,t)\in W$, and we set $q_W=1$ to obtain the corresponding probabilities for locations i and times t outside the window W. Now assuming that we know $\delta_{it}=d_{it}\in\{0,1\}$ for each location i and time t considered, the log-likelihood function is

given by

$$\begin{aligned}
\ell(q_W|\{y_{it}\},\{d_{it}\}) &= \log\left(\left[\prod_{(i,t)\in W}\mathrm{P}(Y_{it}=y_{it},\delta_{it}=d_{it})\right]\left[\prod_{(i,t)\in \overline{W}}\mathrm{P}(Y_{it}=y_{it},\delta_{it}=d_{it})\right]\right)\\
&= \log\left(\prod_{(i,t)\in W}\mathrm{P}(Y_{it}=y_{it}|\delta_{it}=d_{it})\mathrm{P}(\delta_{it}=d_{it})\right)\\
&\quad + \log\left(\prod_{(i,t)\in \overline{W}}\mathrm{P}(Y_{it}=y_{it}|\delta_{it}=d_{it})\mathrm{P}(\delta_{it}=d_{it})\right)\\
&= \sum_{(i,t)\in W}\left[d_{it}\log(p_{it})+(1-d_{it})\log\left((1-p_{it})e^{-q_W\mu_{it}}\frac{(q_W\mu_{it})^{y_{it}}}{y_{it}!}\right)\right]\\
&\quad + \sum_{(i,t)\in \overline{W}}\left[d_{it}\log(p_{it})+(1-d_{it})\log\left((1-p_{it})e^{-\mu_{it}}\frac{\mu_{it}^{y_{it}}}{y_{it}!}\right)\right]\\
&= \sum_{(i,t)\in W}(1-d_{it})\left[y_{it}\log(q_W)-q_W\mu_{it}\right]+c, \qquad (3.2.6)
\end{aligned}$$

where c is some constant with respect to q_W. It follows that the score function is given by

$$\frac{\mathrm{d}}{\mathrm{d}q_W}\ell(q_W|\{y_{it}\},\{d_{it}\}) = \sum_{(i,t)\in W}(1-d_{it})\left(\frac{y_{it}}{q_W}-\mu_{it}\right). \qquad (3.2.7)$$

Recalling that our alternative hypothesis concerns $q_W > 1$, we set the derivative above equal to zero, solve for q_W, and take as our maximum likelihood estimator

$$\hat{q}_W = \max\left\{1, \frac{\sum_{(i,t)\in W} y_{it}(1-d_{it})}{\sum_{(i,t)\in W}\mu_{it}(1-d_{it})}\right\}. \qquad (3.2.8)$$

Since we do not actually observe the δ_{it}'s, we cannot use the above estimator of q_W directly. This is where the EM algorithm comes in handy.

3.2.3 EM Algorithm

For a given space-time window W, let $\hat{q}_W^{(k)}$ be the estimate of q_W in the k'th iteration of the EM algorithm. As usual for the EM algorithm, we define a function Q to be the expectation of the complete data log-likelihood likelihood given the parameters

obtained in the k'th iteration of the algorithm. That is,

$$\begin{aligned} Q(q_W|\hat{q}_W^{(k)}) &\coloneqq \mathrm{E}\left[\ell(q_W|\{y_{it}\},\{\delta_{it}\})\middle|\{y_{it}\};\hat{q}_W^{(k)}\right] \quad \text{(expectation is w.r.t. all } \delta_{it}) \\ &= \mathrm{E}\left[\sum_{(i,t)\in W}(1-\delta_{it})(y_{it}\log(q_W)-q_W\mu_{it}) + \underbrace{c}_{\text{constant w.r.t. } q_W}\middle|\{y_{it}\};\hat{q}_W^{(k)}\right] \\ &= \sum_{(i,t)\in W}\left(1-\mathrm{E}\left[\delta_{it}\middle|\{y_{it}\};\hat{q}_W^{(k)}\right]\right)(y_{it}\log(q_W)-q_W\mu_{it}) + \mathrm{E}\left[c\middle|\{y_{it}\};\hat{q}_W^{(k)}\right]. \end{aligned}$$

Now, note that

$$\begin{aligned} \mathrm{E}[\delta_{it}|\{y_{it}\};\hat{q}_W^{(k)}] &= \mathrm{P}(\delta_{it}=1|Y_{it}=y_{it};\hat{q}_W^{(k)}) \\ &= \frac{\mathrm{P}(Y_{it}=y_{it}|\delta_{it}=1)\mathrm{P}(\delta_{it}=1)\mathbb{1}\{y_{it}=0\}}{\mathrm{P}(Y_{it}=y_{it}|\delta_{it}=1)\mathrm{P}(\delta_{it}=1)+\mathrm{P}(Y_{it}=y_{it}|\delta_{it}=0)\mathrm{P}(\delta_{it}=0)} \\ &= \frac{p_{it}}{p_{it}+(1-p_{it})\exp\left(-\hat{q}^{(k)}\mu_{it}\right)}\mathbb{1}\{y_{it}=0\} \\ &=: \hat{\delta}_{it}^{(k)}, \end{aligned}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. $Q(q_W|\hat{q}_W^{(k)})$ thus becomes

$$Q(q_W|\hat{q}_W^{(k)}) = \sum_{(i,t)\in W}(1-\hat{\delta}^{(k)})(y_{it}\log(q_W)-q_W\mu_{it}) + \mathrm{E}[c|\{y_{it}\};\hat{q}_W^{(k)}]. \tag{3.2.9}$$

In the maximization step of the EM algorithm, we are to maximize Expression (3.2.9) with respect to $q_W > 1$. From Expression (3.2.8) above, we recognize what the required maximum must be. Our EM algorithm can thus be described as follows:

E-step at iteration k: for all pairs $(i,t) \in W$, set

$$\hat{\delta}_{it}^{(k)} = \frac{p_{it}}{p_{it}+(1-p_{it})\exp\left(-\hat{q}_W^{(k)}\mu_{it}\right)}\mathbb{1}\{y_{it}=0\}. \tag{3.2.10}$$

M-step at iteration $k+1$: set

$$\hat{q}_W^{(k+1)} = \max\left\{1, \frac{\sum_{(i,t)\in W} y_{it}\left(1-\hat{\delta}_{it}^{(k)}\right)}{\sum_{(i,t)\in W}\mu_{it}\left(1-\hat{\delta}_{it}^{(k)}\right)}\right\}. \tag{3.2.11}$$

This procedure is repeated until convergence, for each window $W \in \mathcal{W}$. Cançado et al. (2014) used the initial value $\hat{\delta}_{it}^{(0)} = 0.5$ if the corresponding count was 0 and $\hat{\delta}_{it}^{(0)} = 0$ otherwise, and used the absolute convergence criterion $|\hat{\delta}_{it}^{(k+1)} - \hat{\delta}_{it}^{(k)}| < 0.01$.

In difference to the method in Cançado et al. (2014), the estimates $\hat{\delta}_{it}^{(k)}$ for the location-time pairs (i,t) outside the space-time window W do not need to be calculated, as they are irrelevant for inference about q_W. In reference to the hypotheses defined above, it is clear that these estimates will be the same outside of W. However, the estimates do need to be calculated under the null hypothesis (when $q_W = 1$) inside of W; for each $(i,t) \in W$ we set

$$\delta_{it}^{\dagger} = \frac{p_{it}}{p_{it} + (1 - p_{it}) \exp(-\mu_{it})} \mathbb{1}\{y_{it} = 0\}. \tag{3.2.12}$$

3.2.4 The EB-ZIP Scan Statistic

The above EM estimation above is repeated for each potential cluster W, arriving at final estimates q_W^* and $\delta_{it}^* = \delta_{it,W}^*$ on convergence of the algorithm, along with the estimates $\delta_{it}^{\dagger} = \delta_{it,W}^{\dagger}$. Then, as in Cançado et al. (2014), we use the full likelihood of Equation (3.2.6) to form the ratio of likelihoods

$$\lambda_W = \frac{L(q_W^* | \{y_{it}\}, \{\delta_{it}^*\})}{L(1 | \{y_{it}\}, \delta_{it}^{\dagger})} \tag{3.2.13}$$

$$= \frac{\prod_{(i,t)\in W} p_{it}^{\delta_{it}^*} \left[(1 - p_{it}) e^{-q_W^* \mu_{it}} \frac{(q_W^* \mu_{it})^{y_{it}}}{y_{it}!} \right]^{1-\delta_{it}^*}}{\prod_{(i,t)\in W} p_{it}^{\delta_{it}^{\dagger}} \left[(1 - p_{it}) e^{-\mu_{it}} \frac{\mu_{it}^{y_{it}}}{y_{it}!} \right]^{1-\delta_{it}^{\dagger}}}. \tag{3.2.14}$$

The expectation-based ZIP scan statistic (EB-ZIP) is then given by

$$\lambda^* = \max_W \lambda_W, \tag{3.2.15}$$

and the most likely (space-time) cluster W^* for a disease outbreak is given by the W that corresponds to this maximum.

3.3 Software

The R programming language and software environment was used to implement the methods described in this thesis, as well as carry out the simulations described in Chapter 5 and Chapter 4. The methods are available as the R package `scanstatistics`, accessible from the GitHub repo https://github.com/BenjaK/scanstatistics. It can be installed using the R command

```
devtools::install_github("benjak/scanstatistics")
```

This R package supplements existing packages for the monitoring of count time series

and outbreak detection, such as the `surveillance` package (Höhle et al., 2015), whose capabilities are described in Höhle (2007) and Salmon et al. (2015). The books *R Packages* and *Advanced R* by Wickham (2014, 2015) were immensely helpful in the design of this package. Also of immense use was the package `data.table` (Dowle et al., 2014).

4 Simulation Study

The expectation-based ZIP scan statistic is the main contribution of this thesis, and interest therefore lies in measuring its performance on the type of data and outbreaks it was designed for—spatial time series of zero-inflated Poisson counts with hot-spot outbreaks. If it does not perform well on that type of data in a controlled setting, it reasonably stands little chance of detecting outbreaks on data from the real world. We thus begin by comparing the proposed expectation-based zero-inflated Poisson scan statistic against the expectation-based Poisson scan statistic (Neill et al., 2005) and the hot-spot efficient score scan statistic (Tango et al., 2011), on counts randomly generated from a zero-inflated Poisson distribution. For brevity, we refer to these scan statistics simply as ZIP, PO, and NB, names corresponding to the distributions on which they are based. Each of these three scan statistics was designed for hot-spot outbreaks (see Section 2.2.3), so this is the type of outbreak that will be simulated. Because Tango, Takahashi & Kohriyama's scan statistic for 'emerging outbreaks' (as described in Section 3.1.2) was not designed for this type of outbreak, it seems less relevant and is thus excluded from the comparison. It will be used in the next chapter, however. We begin by describing the design of this simulation study and how the comparison between the different scan statistics will be made.

4.1 Design

For our simulation study we consider a square 14×14 grid of points, each point representing a single spatial region, and its coordinates the centroid of the region. This grid—our study area—is depicted in Figure 4.1. Because only the centroids are important for distance calculations in the simulations we run, the boundaries of the regions are of less importance, and are not drawn in the figure.

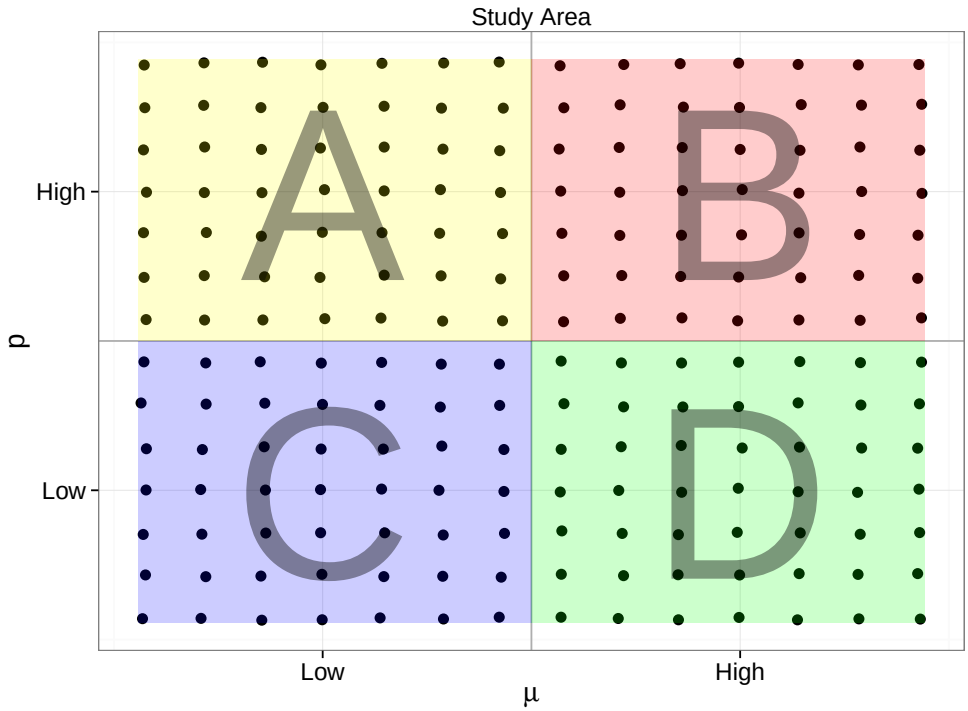

Figure 4.1 Grid of points, representing the study area. The study area is divided into four subareas, corresponding to different ranges of the parameter values for the ZIP distributions from which the simulated data is generated.

As can be seen, the study area is divided into four subareas denoted A, B, C, and D. For each region i, a baseline set corresponding to 15 weeks of non-outbreak data is generated from a zero-inflated Poisson distribution with parameters p_i and μ_i. To put focus on the actual scan statistics rather than the parameter estimation methods, we suppose that distribution parameters are constant over time and only differ between regions. The parameter μ_i of the counts in subareas A and C are given low values, generated uniformly at random from the range $[0.5, 3.5]$. In subareas B and D, the μ_i's are comparatively high, similarly generated in the range $[15, 30]$. Likewise, the excess zero probabilities p_i are low in subareas C and D ($p_i \in [0.05, 0.15]$), and high in subareas A and B ($p_i \in [0.5, 0.65]$). As was done in the simulation study conducted by Tango et al. (2011) the parameter values and baseline counts generated from their use are drawn *once*, as computation times would be prohibitive

if the simulation results were averaged over multiple sets of baseline parameter values and counts. The choice of ranges for these parameters is motivated by values seen when fitting a zero-inflated Poisson distribution to district-level disease data (Salmonellosis Enteritidis) from Germany, obtained from the Robert Koch Institute (Robert Koch Institute: SurvStat@RKI 2.0, 2015). The idea behind the choice of high or low parameter values is to compare the PO, NB, and ZIP scan statistics on outbreaks in different settings.

In Figure 4.2, relative frequency histograms are shown for the baseline counts in each of the subareas A, B, C, and D. The counts in each area are considered jointly over all 144 regions and 15 time points, since the parameter values of the distributions that generated the counts are similar in a given area.

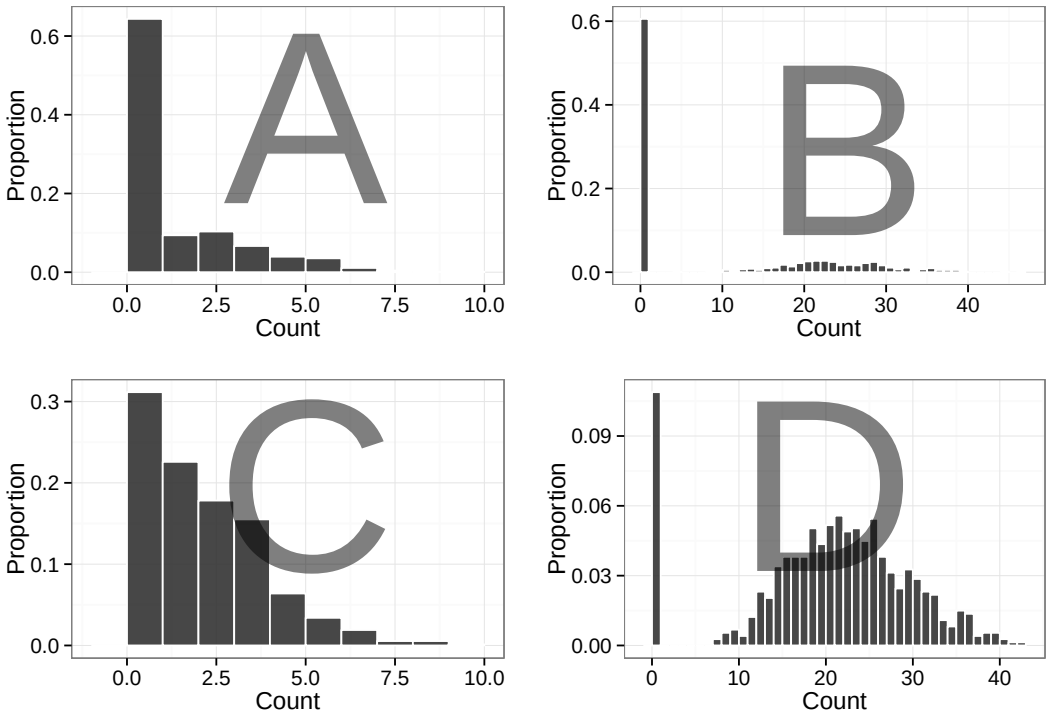

Figure 4.2 Relative frequency histograms of the baseline (non-outbreak) dataset, one histogram for all the counts in each of the subareas A, B, C, and D.

On the baseline data set, for each region, the expected value parameter μ_i for the Poisson and negative binomial are estimated using the region-specific sample mean $\bar{y}_i$. The parameter ϕ_i of the negative binomial distribution is estimated using the moment estimator

$$\hat{\phi}_i = \begin{cases} \frac{\bar{y}_i^2}{s_i^2 - \bar{y}_i}, & \text{if } s_i^2 > \bar{y}_i \\ \infty, & \text{otherwise,} \end{cases} \tag{4.1.1}$$

as was used by Tango et al. (2011, p. 108). s_i^2 is here the sample variance for the counts in region i. Because the method of moments may yield a negative estimate for

the excess zero probability p_i when used with the ZIP distribution, the parameters of the ZIP distribution are estimated using the EM algorithm instead. The derivation of the EM algorithm for the ZIP distribution is given in Appendix A.

Outbreaks will be generated in a subset of regions in each of the subareas A, B, C, and D, and these outbreaks will vary in severity q—the factor by which the Poisson parameter is increased in an outbreak—and duration T. To exemplify, suppose an outbreak is to be simulated in subarea A. We first choose an outbreak zone Z_{A}, consisting of a number of regions in A. We also choose a duration of the outbreak, say $T = 3$ weeks, and an outbreak severity, say $q = 2$. For each region $i \in Z_{\mathrm{A}}$, we then generate 3 counts from a $\mathrm{ZIP}(p_i, 2 \cdot \mu_i)$ distribution, and for all regions not in Z_{A}, 3 counts are generated from a $\mathrm{ZIP}(p_i, \mu_i)$ distribution. The parameters p_i and μ_i are the same as for the baseline data set. Then, using the parameters estimated on the baseline data set for each distribution (Poisson, negative binomial, ZIP), we calculate the corresponding scan statistics and their most likely cluster. The value of each of these scan statistics is then compared to 999 Monte Carlo replicates, as described Section 2.2.5. The replicates are generated using the estimated parameters. The calculated p-value is compared to the significance level 0.02, corresponding to approximately one false alarm every year (and the same value chosen by Tango et al., 2011). For a p-value lower than the significance level, the spatial component of the detected most likely cluster (MLC) Z^* can then be compared to the true cluster Z_{A}, as described next. In Figure 4.3, four examples of simulated outbreaks—one in each of the subareas A, B, C, and D—are shown. The counts in the first 15 weeks, highlighted in green, shows the time series of baseline counts for each region plotted in the same graph. In the last 3 weeks, highlighted in red, counts in five regions (5 nearest neighbors) of each subarea are generated from ZIP distributions for which the Poisson parameters are 2 times what they were in the baseline period. In the remaining regions, the counts are generated from distributions with the same parameters as in the baseline period.

In our simulations, the true outbreak zones Z_j, $j = A, B, C, D$, will always be the corner regions and their 4 nearest neighbors. Three different outbreak severities $q \in \{1.5, 2, 2.5\}$ are considered, as well as two outbreak durations $T \in \{1, 3\}$ weeks, if we work with time units of one week. For outbreaks with duration 3, the analysis is done in the third week. For each combination of true outbreak zone Z_j, severity q, and duration T, 1000 outbreaks are simulated, and the PO, NB, and ZIP scan statistics calculated along with their respective MLCs (spatial component only). The set $\mathcal{Z}$ of all potential spatial outbreak zones Z is taken as the set of all k nearest neighbors of each point (region) on the grid, for $0 \leq k \leq 9$. Thus, the zones Z considered consist of between 1–10 regions. Before calculating (Euclidean) distances between points on the grid, a small noise is added in order to prevent ties.

In evaluating the results of the simulation study, we focus mainly on the spatial detection accuracy of the scan statistics, not their ability to detect the true duration of outbreaks. Rather, we look at how inclusion of the temporal dimension in the

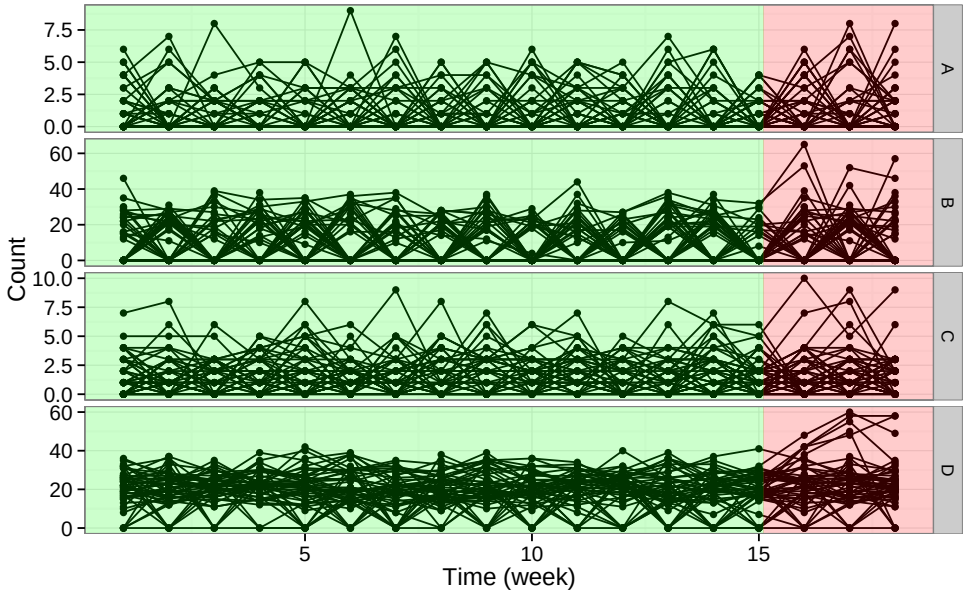

Figure 4.3 Overlapping time series for all 196 regions in the study area. The first 15 weeks show counts during the baseline period; the last 3 weeks show outbreaks generated in 5 regions of each subarea A, B, C, and D.

calculation of the scan statistics can aid in detecting the true outbreak regions. If a simulated outbreak is detected by one of the scan statistics—meaning that the p-value for the statistic is lower than our chosen significance level—we wish to compare the detected most likely (spatial) cluster Z^* to the true spatial outbreak cluster Z^{true}. Neill (2009*b*, p. 506) proposes two primary measures that can be used for this purpose: the *spatial precision* and the *spatial recall*. These measures take values in the range $[0, 1]$, with values closer to 1 being better. The spatial precision (also known as *positive predicted value*) is the proportion of regions detected by our method that are true outbreak regions, and is calculated as

$$\text{Spatial precision} = \frac{|Z^* \cap Z^{\text{true}}|}{|Z^*|} \tag{4.1.2}$$

The spatial precision gives us a measure of how relevant the outbreak zone (MLC) we found is, in terms of its size and how many true outbreak regions are in it. A large and significant MLC containing all true outbreak regions but also many other non-outbreak regions will have a low precision, while a significant MLC with only one region—which happens to be a true outbreak region—will have a perfect precision of 1.

The spatial recall (or *sensitivity*) is similarly the proportion of true outbreak regions that are detected by our method, and is calculated as

$$\text{Spatial recall} = \frac{|Z^* \cap Z^{\text{true}}|}{|Z^{\text{true}}|} \tag{4.1.3}$$

A high spatial recall may not necessarily be good: if the detected cluster Z^* consists of all regions in the study area the recall will be 1, but for an outbreak affecting only a few regions the cluster Z^* won't actually tell us anything about where the true outbreak cluster might be.

To see that the precision and recall measures can differ substantially, consider an outbreak that affects 15 regions, of which our scan statistic detects 3, plus one region that is not actually affected. In this case, the spatial precision is $3/4 = 0.75$ while the spatial recall is only $3/15 = 0.2$. As an aggregate measure of a scan statistic's spatial detection accuracy, Neill (2009*b*, p. 506) uses the *F-score*, which is the harmonic mean of the precision and recall:

$$F = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.1.4}$$

$$= 2 \cdot \frac{|Z^* \cap Z^{\text{true}}|}{|Z^*| + |Z^{\text{true}}|} \tag{4.1.5}$$

These three measures will be calculated for the significant clusters found by each scan statistic, for each combination of outbreak severity, duration, and affected zone.

We also compare the false positive ratios of the PO, NB, and ZIP scan statistics by generating three ZIP-distributed counts for each region using the true baseline parameters p_i and μ_i, 1000 times over. On each such simulation of a non-outbreak period of 3 weeks, each scan statistic is calculated and a p-value calculated using the Monte Carlo replicates mentioned earlier. If this p-value is lower than the significance level 0.02, the null hypothesis of no outbreak is (falsely) rejected. Since the scan statistics were calculated on data generated from parameters believed to hold under the null hypothesis of no outbreak, we would hope that the proportion of rejected null hypotheses would match our significance level. The false positive ratio is given by this proportion.

4.2 Results

The false positive ratios calculated for the 1000 simulated null hypothesis data sets are 100%, 7.5%, and 10.4% for the PO, NB, and ZIP scan statistics, respectively. The expectation-based Poisson scan statistic clearly performs worst, raising a false alarm 100% of the time. This is due to the fact that the Poisson scan statistic uses the sample means calculated for each region in the baseline data as parameters. Due to the presence of excess zeros, these sample means are likely to be low in comparison to the Poisson parameter in the ZIP distribution which generated the baseline data and the simulated null hypothesis data sets on which the false positive ratio was calculated. When the Monte Carlo p-value is calculated for the Poisson scan statistic, it is based on scan statistics calculated on data that has been generated using the

low sample means as parameters. These scan statistics will be small in comparison to the 'observed' scan statistics, as these are calculated on data generated from ZIP distributions whose Poisson parameters are comparatively high. On the other hand, the NB and ZIP scan statistics raise false alarms at much lower rates, but still above the significance level 0.02. For the ZIP scan statistic at least, the false positive ratio should improve with better parameter estimates.

For the simulated outbreaks, we begin by showing the spatial precision, recall, and F-score for the detected outbreaks, i.e. those whose corresponding scan statistic had a p-value below 0.02. In Figure 4.4, we show the spatial precision of the three scan statistics for the different outbreak scenarios with severity $q = 1.5$, which again is the factor by which the Poisson mean μ_i of the ZIP-distribution is multiplied by for a region i affected by the outbreak.

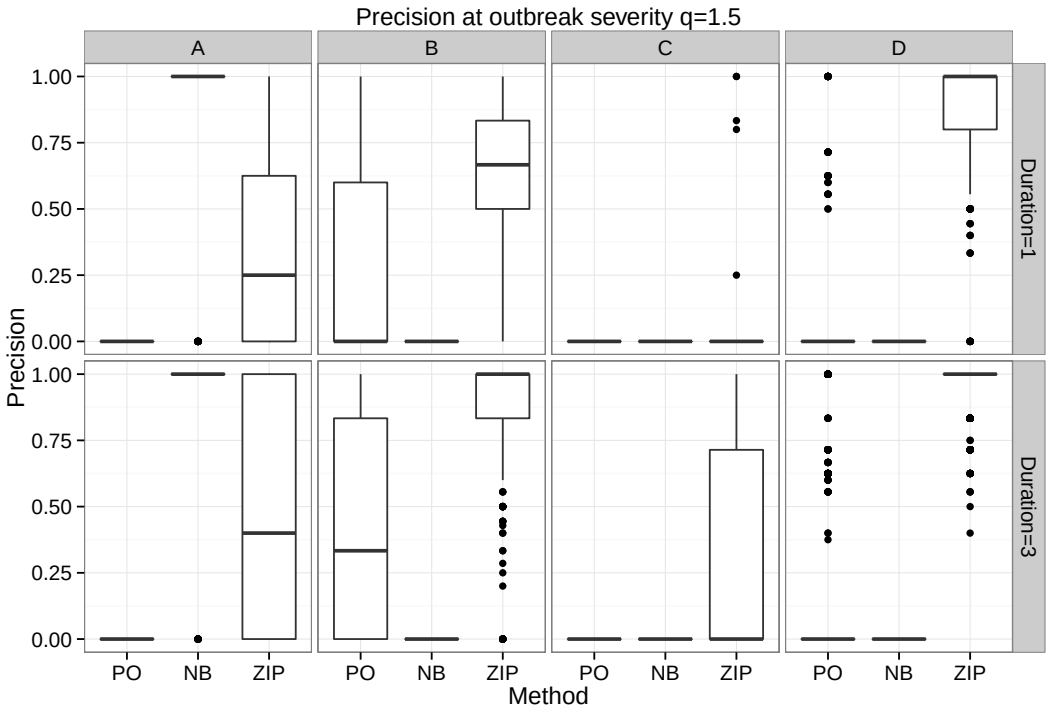

Figure 4.4 Boxplot of the spatial precision for the three scan statistics calculated for the 1000 simulated outbreaks, for the scenarios in which the outbreak severity q is set at 1.5.

Figure 4.4 shows that the ZIP scan statistic generally performs much better than the PO and NB scan statistics, having a high proportion of true outbreak regions among all regions detected. The exception is outbreaks in area A, in which the ZIP parameter μ_i is low and the parameter p_i is high. Here, the NB scan statistic has an almost perfect precision, meaning that all outbreak regions detected are also true outbreak regions. Also of interest is that the ZIP scan statistic seems to have a higher precision for outbreaks in regions B and D, in which the ZIP parameter μ_i is high. Lastly, the precision for the ZIP scan statistic generally seems higher for

outbreaks of duration 3 than those of duration 1, indicating that the method gains accuracy when considering longer outbreak durations.

Figure 4.5 similarly shows the spatial recall of the three methods. Here, the ZIP scan statistic performs best on all of the areas A, B, C, and D, and again seems to do better in areas B and D, finding a higher proportion of true outbreak regions than it does for outbreaks in areas A or C. The recall for the ZIP statistic is also better for outbreaks of duration 3 compared to those of duration 1, similar to the spatial precision.

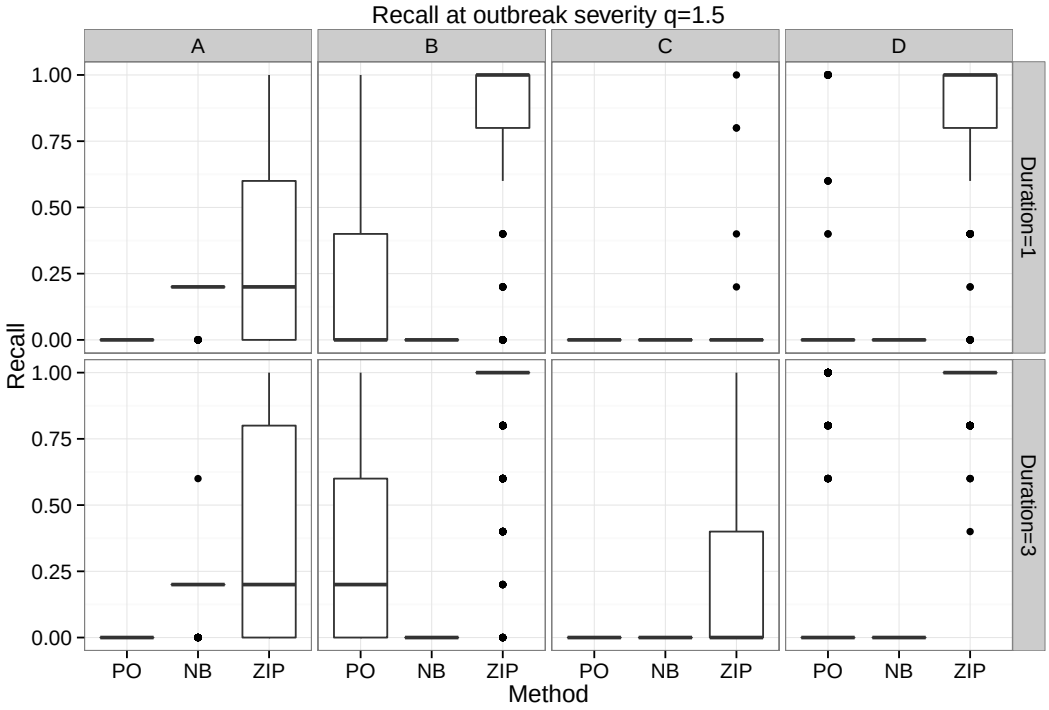

Figure 4.5 Boxplot of the spatial recall for the three scan statistics calculated for the 1000 simulated outbreaks, for the scenarios in which the outbreak severity q is set at 1.5.

Figure 4.5 also shows that all three methods struggle to detect true outbreaks that occur in area C, which was also evident from the precision in Figure 4.4. In area C, both the excess zero probability and the Poisson mean parameter of the ZIP distribution are low, and increasing the mean by a factor of 1.5 is apparently insufficient to cause detectable outbreaks.

To summarize the two previous figures, Figure 4.6 shows boxplots of the F-score, which again is the harmonic mean of the spatial precision and spatial recall.

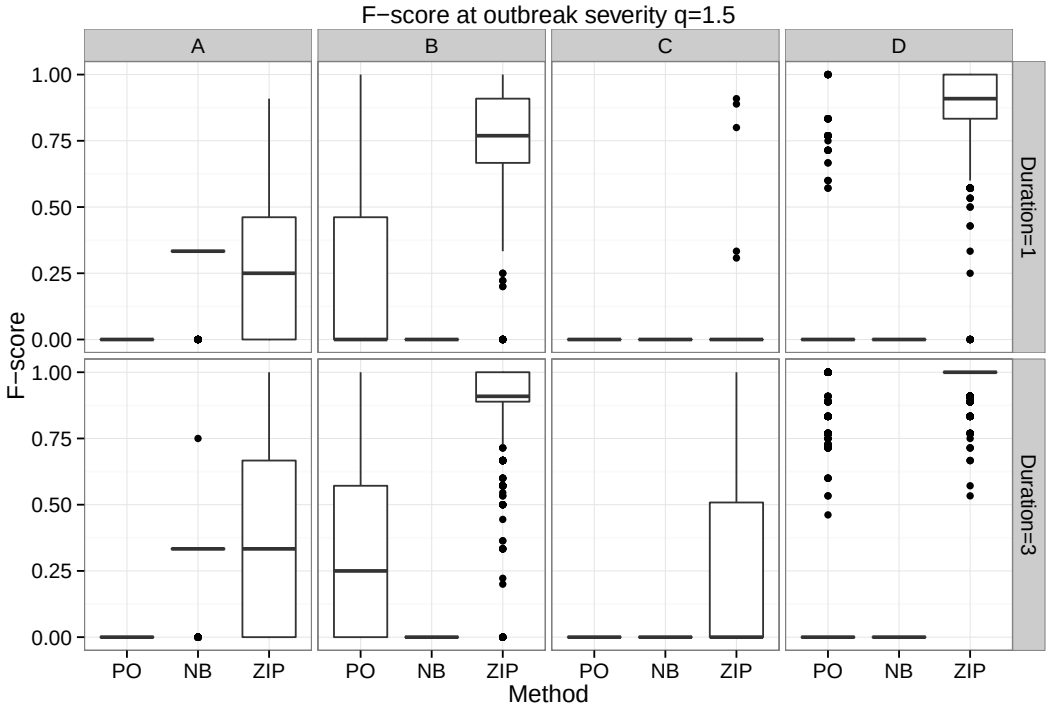

Figure 4.6 Boxplot of the F-score for the three scan statistics calculated for the 1000 simulated outbreaks, for the scenarios in which the outbreak severity q is set at 1.5.

Though Figure 4.6 holds no surprises, it affirms the statement that the spatial detection accuracy of the proposed ZIP scan statistic is superior to the PO and NB scan statistics, when data behaves according to the outbreak model for which the ZIP scan statistic was designed. This indicates that the ZIP scan statistic may be of some value in the real world, when data is though to resemble that generated by a zero-inflated Poisson process.

Boxplots corresponding to those above but for simulated outbreaks with severities $q \in \{2, 2.5\}$ can be found in Appendix B. Though the ZIP scan statistic still performs best in these scenarios, the performance of the PO and NB scan statistic is improved for higher values of the outbreak severity. This is reasonable, as any scan statistic ought to detect an outbreak of sufficiently large magnitude.

It is also interesting to see what number of detected outbreaks that the plots above are based on. In Table 4.1, the number of detected outbreaks per 1000 simulations are given for each scan statistic and outbreak scenario.

Table 4.1 Number of significant clusters detected by each scan statistic, in 1000 simulated outbreaks at the specified area (A, B, C, or D) and with outbreak duration $T \in \{1, 3\}$ and severity $q \in \{1.5, 2, 2.5\}$ as specified by the leftmost column.

	PO				NB				ZIP			
Severity	A	B	C	D	A	B	C	D	A	B	C	D
Duration = 1												
1.5	1000	1000	1000	1000	101	36	32	35	41	295	22	796
2	1000	1000	1000	1000	185	44	26	45	145	842	90	999
2.5	1000	1000	1000	1000	285	34	47	600	271	922	249	1000
Duration = 3												
1.5	1000	1000	1000	1000	170	71	66	65	248	914	164	1000
2	1000	1000	1000	1000	322	71	80	759	565	996	570	1000
2.5	1000	1000	1000	1000	510	196	255	993	828	1000	916	1000

In reference to the earlier plots of spatial precision and recall, Table 4.1 shows that even though the PO scan statistic always detects an outbreak in our simulations—which by itself could be viewed as a good thing—the outbreak zones detected correspond very little to the true outbreak zones. Thus, yet again, the PO scan statistic gives false alarms most of the time, for the same reasons given in the beginning of this section. On the other hand, the NB scan statistic performed better in terms of spatial accuracy, but the number of detected outbreaks is far from a full 1000, particularly at lower outbreak severities. In comparison, the ZIP scan statistic detects more (and more relevant) outbreaks, but the performance is again worse for simulated outbreaks in areas A and C compared to those in areas B and D. The performance improves at higher levels of the outbreak severity, however.

The conclusion of the simulation study in this chapter is that the proposed expectation-based ZIP scan statistic may hold some promise, outperforming comparable methods in terms of outbreak detection ability on data and outbreaks simulated according to the models that the method was designed for. A second conclusion is that inclusion of the time dimension when detecting outbreaks is valuable, as it seems to increase both detection rate (seen by the number of significant outbreaks detected per 1000 simulated outbreaks) and spatial detection accuracy.

5 Case Study: Cryptosporidiosis Outbreak in Germany

The previous chapter proved by simulation that the proposed expectation-based ZIP scan statistic is capable of accurately detecting outbreaks on the type of outbreak data it was designed for. It was moreover shown that its performance was in most cases superior to two other expectation-based scan statistics: the Poisson scan statistic proposed by Neill et al. (2005) and the negative binomial score scan statistic proposed by Tango et al. (2011) In this chapter, we return to the outbreak of cryptosporidiosis in the city of Halle, Germany, that was discussed in the introduction (Chapter 1). This data, obtained from the SurvStat database that is maintained by the Robert Koch Institute (RKI) in Germany (Robert Koch Institute: SurvStat@RKI 2.0, 2015) and which was plotted in Figure 1.2, is shown again in Figure 5.1.

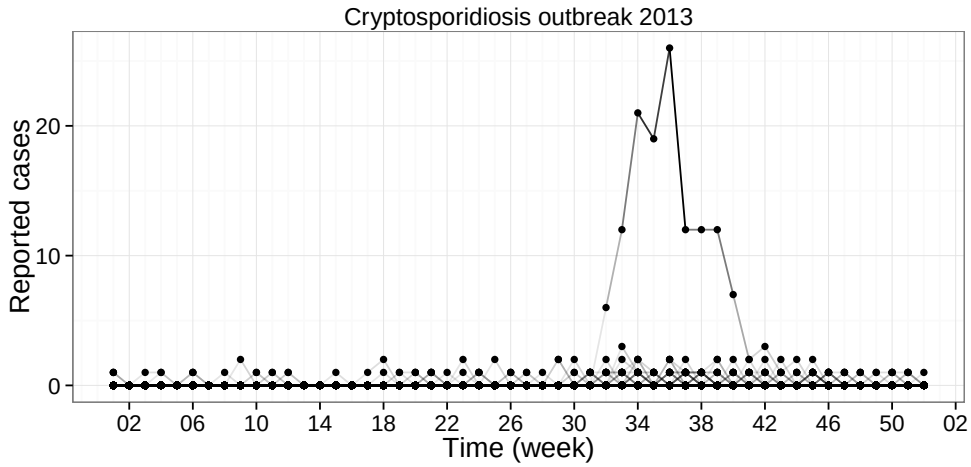

Figure 5.1 Time series of reported cryptosporidiosis cases for all 402 districts of Germany; each line is a time series for a single district, but most overlap at zero at all time points. The clear outlying line shows the reported cases for the city of Halle (Saale).

Though this outbreak data is not ideal for the outbreak detection problem considered in this thesis—it affects just a single region, rather than multiple neighboring regions—it can still tell us something about the relative merits of the scan statistics

examined here. Given the high frequency of zero counts in the data, it is particularly interesting to see how the proposed expectation-based ZIP scan statistic will perform. In this chapter, we also consider the 'emerging outbreak' scan statistic of Tango et al. (2011) that was covered in Section 3.1.2, but not included in the simulation study of the previous chapter. According to Gertler et al. (2015), the outbreak in Halle was detected by the local health department in that city, and at federal level by the Robert Koch Institute in Berlin in week 32. In two weeks prior, the case counts of cryptosporidiosis was zero in the city of Halle. We therefore run our analysis in weeks 32–34: the analysis run in week 32 considers a maximum outbreak duration of 1, in week 33 a maximum duration of 2, and in week 34 a maximum duration of 3. The per-district counts for the first 31 weeks of the year are used to estimate the parameters of the Poisson, negative binomial, and zero-inflated Poisson distributions corresponding to our four scan statistics. In this baseline period, case counts of cryptosporidiosis are zero in all 31 weeks for all but 34 of the districts of Germany, and counts are never higher than 2 in any one week. It thus seems justified to fit parsimonious models in which temporal covariates are not included. In reference to section 2.2.4, we fit three generalized linear mixed models with the following structures for the conditional expected values:

$$\text{PO}: \quad \log(\mu_{it}) = \lambda + a_i \tag{5.0.1}$$

$$\text{NB}: \begin{cases} \log(\mu_{it}) = \gamma + b_i \\ \phi_{it} = \phi \text{ (constant)} \end{cases} \tag{5.0.2}$$

$$\text{ZIP}: \begin{cases} \log(\mu_{it}) = \zeta + c_i \\ \text{logit}(p_{it}) = \rho + d_i, \end{cases} \tag{5.0.3}$$

where $\text{logit}(p) = \log(p/(1-p))$ for $p \in (0,1)$, and $a_i \sim N(0, \sigma_a^2), b_i \sim N(0, \sigma_b^2), c_i \sim N(0, \sigma_c^2)$ and $d_i \sim N(0, \sigma_d^2)$ are the random district-specific intercepts conditioned on. The models were fitted using the R package `lme4` (Bates et al., 2015*a*,*b*), though a custom routine had to be written to fit the ZIP model as shown above. This routine alternates between fitting a logistic mixed effects model for the zero and non-zero counts, and a weighted Poisson GLMM for the Poisson parameter of the ZIP distribution. The program is supplied in Appendix C. In Table 5.1, the parameter estimates for the above models are shown.

Table 5.1 Parameter estimates for the fitted mixed models.

	PO		NB			ZIP			
Parameter	λ	σ_a	γ	ϕ	σ_b	ζ	ρ	σ_c	σ_d
Estimate	−10.06	5.5	−10.02	0.05	1.5	−10.13	10.03	5.6	5.4

Fitted values for the parameter μ_{it} for each model are very close to zero for all districts, a large majority of values of the order 10^{-4} or lower due to the many regions with all zeros during the baseline period. The excess zero probabilities p_{it} are close to 1, though this should not be interpreted as a massive underreporting

of cases or similar, since the number of reported cases are expected to be very low as well. The estimate of the negative binomial parameter ϕ is approximately 0.05, which is natural if recalling that the variance of the negative binomial distribution is $\mu_i + \mu_i^2/\phi$. Because the mean parameter μ_{it} is so small, the estimate of ϕ has to be small as well in order to make the variance account for all non-zero values in the data. For the spatial zones over which we scan, we take the k nearest neighbors of each district in Germany, for $0 \leq k \leq 9$. That is, the maximum cluster size is 10. Because we noticed a tendency of the ZIP scan statistic to overestimate the size of the true outbreak cluster, we also considered maximum cluster sizes of 5, 2, and 1, for comparison. Pairwise distances were calculated between the geographic centroids of each district using the great circle distance (WGS84 ellipsoid), as implemented in the R package `sp` (Bivand et al., 2013; Pebesma & Bivand, 2005). To determine significance of the highest-scoring clusters found in our analysis, we will for each scan statistic calculate a p-value using Equation (2.2.13), based on 999 Monte Carlo replicates of the statistic.

The results of the analysis are shown in Table 5.2 on the next page. In this table, we again refer to the Poisson scan statistic as PO and the zero-inflated Poisson scan statistic as ZIP. The negative binomial score scan statistic based on a hot-spot outbreak model, as given in Section 3.1.1, is referred to as NB-HS. The 'emerging outbreak' negative binomial score scan statistic of Section 3.1.2 is referred to as NB-EM. The first result from running the analysis for the three scan statistics is that the negative binomial score scan statistic fails to detect the true outbreak district. This is true for all maximum cluster sizes considered, and all weeks of analysis. This may be explained by referring to Equation (3.1.9) and the equations prior: the parameter μ_i is relatively large in the true outbreak district (city of Halle), since it was one of the few districts which had seen non-zero counts in the baseline period. This makes the overdispersion parameter $w_i = 1 + \mu_i/\phi$ large as well, which decreases the size of the numerator in Equation (3.1.9) relative to what would be its size with a lower value of w_i. On the contrary, there are other districts with much lower means, which decreases the size of the numerator in Equation (3.1.9), thus blowing up the value of the scan statistic. These are the districts detected by the negative binomial score scan statistic, even though the maximum of the counts in them are only equal to 1.

Table 5.2 Results from analysis of cryptosporidiosis outbreak in Halle.

Method	Analysis week	Maximum cluster size	p-value	Halle detected	MLC size
PO	32	10	0.001	TRUE	3
PO	32	5	0.001	TRUE	3
PO	32	2	0.001	TRUE	1
PO	32	1	0.001	TRUE	1
PO	33	10	0.001	TRUE	10
PO	33	5	0.001	TRUE	3
PO	33	2	0.001	TRUE	1
PO	33	1	0.001	TRUE	1
PO	34	10	0.001	TRUE	10
PO	34	5	0.001	TRUE	3
PO	34	2	0.001	TRUE	1
PO	34	1	0.001	TRUE	1
NB-HS	32	10	0.001	FALSE	1
NB-HS	32	5	0.001	FALSE	1
NB-HS	32	2	0.001	FALSE	1
NB-HS	32	1	0.001	FALSE	1
NB-HS	33	10	0.001	FALSE	1
NB-HS	33	5	0.001	FALSE	1
NB-HS	33	2	0.001	FALSE	1
NB-HS	33	1	0.001	FALSE	1
NB-HS	34	10	0.001	FALSE	1
NB-HS	34	5	0.001	FALSE	1
NB-HS	34	2	0.001	FALSE	1
NB-HS	34	1	0.001	FALSE	1
NB-EM	32	10	0.001	FALSE	1
NB-EM	32	5	0.001	FALSE	1
NB-EM	32	2	0.001	FALSE	1
NB-EM	32	1	0.001	FALSE	1
NB-EM	33	10	0.001	FALSE	1
NB-EM	33	5	0.001	FALSE	1
NB-EM	33	2	0.001	FALSE	1
NB-EM	33	1	0.001	FALSE	1
NB-EM	34	10	0.001	FALSE	1
NB-EM	34	5	0.001	FALSE	1
NB-EM	34	2	0.001	FALSE	1
NB-EM	34	1	0.001	FALSE	1
ZIP	32	10	0.001	TRUE	10
ZIP	32	5	0.001	TRUE	5
ZIP	32	2	0.002	FALSE	2
ZIP	32	1	0.002	FALSE	1
ZIP	33	10	0.001	FALSE	10
ZIP	33	5	0.001	FALSE	5
ZIP	33	2	0.001	TRUE	2
ZIP	33	1	0.001	TRUE	1
ZIP	34	10	0.001	TRUE	10
ZIP	34	5	0.001	TRUE	5
ZIP	34	2	0.001	TRUE	2
ZIP	34	1	0.001	TRUE	1

Focusing first on the analyses run in week 32, the second result seen in Table 5.2 is that both the Poisson and ZIP expectation-based scan statistics manage to detect the true outbreak district Halle, though the ZIP statistic reports as the most likely cluster (MLC) Halle and its 9 nearest neighbors. Even though the counts in all but one of the true outbreak district's nearest neighbors have case counts of 0 in week 32, they can still contribute to the value of the scan statistic through the quantity δ_{it}^*, as defined in Section 3.2.4. In reference to the spatial precision measure used in Chapter 4, this is obviously a drawback of the proposed ZIP-scan statistic. One may reason however, that if underreporting is prevalent, regions with zero counts near a region with a relatively high number of counts should still be monitored, if there is reason to believe that the disease could or already has spread to nearby areas. In comparison, the expectation-based Poisson scan statistic only reports as the MLC the true outbreak region and its two closest neighbors, which is better but still not a perfect result.

Because of these results, we reran the analysis for potential clusters Z of maximal size 5, 2, and 1. In these analyses, the Poisson scan statistic continues to do well, but for the expectation-based ZIP scan statistic, the true outbreak district is no longer in the most likely cluster when the maximal cluster size is two or one when the analysis is run in week 32. However, Halle still registers as a secondary significant cluster, as the p-value calculated for the statistic of it is still below 0.02, which was the significance level used in the previous chapter. If the analysis for the ZIP scan statistic is run in week 34 instead, the method manages to detect Halle as an outbreak region for all maximum cluster sizes. For the analysis run in week 33, this is only true when the maximum cluster size is either 1 or 2. One may also note that the ZIP statistic always reports a cluster of maximum possible size. On the other hand, the two negative binomial score scan statistics always seem to report an MLC of maximum size 1. In weeks 33 and 34, the Poisson scan statistic also shows a tendency to report a large cluster size.

In conclusion, this chapter has applied the expectation based Poisson, negative binomial score, and ZIP scan statistics on real-world data from an outbreak of cryptosporidiosis. The results indicate that Poisson and zero-inflated Poisson scan statistics have some potential to detect outbreaks in this type of data, which is characterized by an abundance of zero counts and few weekly case counts above two. However, both methods tend to overestimate the true cluster size, the ZIP scan statistic more so than the Poisson. The two negative binomial score scan statistics fail completely in detecting the true outbreak region. It would be interesting to do the same comparison on a different set of real outbreak data, perhaps with higher average weekly counts in non-outbreak conditions, and with an outbreak that affected more than one region. Such knowledge about actual outbreak locations is scarce however, and the above comparison will have to do for this thesis. We now conclude this thesis by discussing what has been learned in this chapter and the previous ones.

6 Summary and Discussion

The main purpose of this thesis has been to present a novel method for detecting disease outbreak clusters in a spatiotemporal setting, using past observations gathered from a multitude of geographical regions to inform the analysis of current data. This novel method is a scan statistic based on the zero-inflated Poisson (ZIP) distribution, which draws inspiration from a similar scan statistic proposed in an article by Cançado et al. (2014). The ZIP distribution is appropriate for outbreak detection in situations when some local health centers lack the facilities to diagnose a given disease or when reported counts are biased downwards; the latter could be due to e.g. underreporting or the lack of access to medical care for uninsured individuals. After establishing the general scan statistics methodology in Chapter 2, the derivation of the proposed ZIP scan statistic was presented in Chapter 3, along with another recent scan statistic formulated by Tango et al. (2011). Tango, Takahashi & Kohriyama's scan statistic, based on a negative binomial distribution, was intended to supplement the the Poisson scan statistic presented in Chapter 2 as a method that could better handle overdispersed data. These two methods were then used for comparison with the proposed ZIP scan statistic, first on simulated data in Chapter 4, then in a case study on real-world outbreak data in the case study of Chapter 5.

The simulation study in Chapter 4 illustrated that the proposed ZIP scan statistic holds some promise on well-behaved data, in the sense that its spatial detection accuracy was higher than those of the Poisson and negative binomial scan statistics when outbreaks were simulated in time series of counts drawn from a zero-inflated Poisson distribution. The performance advantage was particularly evident when the Poisson parameter of the ZIP distribution was high, but less clear when the magnitude of the outbreak was higher, as illustrated by the plots in Appendix B. The simulations also showed that inclusion of the time dimension in the analysis improved the detection accuracy, though the goal is of course to detect outbreaks as soon as possible. If an outbreak can be detected in week one rather waiting for data to accumulate until week three, that is better.

Finally, the case study in Chapter 5 showed that proposed ZIP scan statistic had a deficiency in that it detected a much larger outbreak cluster than what really was the case, due to the inclusion of regions with counts of zero in the detected cluster. This apparent flaw is inherent in the mathematical construction of the statistic, but could perhaps be remedied in an ad hoc manner by focusing attention on those regions in the detected cluster that have positive counts, until a more rigorous procedure

has been devised. Another drawback of the proposed scan statistic that cannot be completely circumvented is computation time. In one experiment performed, the run time of the ZIP scan statistic was nearly 50 times greater than that of the simple Poisson scan statistic. This is due to the sometimes many iterations of the EM algorithm that are required, which increase with the number of potential outbreak zones considered. The computation times could possibly be decreased by choosing less strict convergence criteria. However, the algorithms implemented for the analyses in this paper were written exclusively in R, and significant speedups may be achievable by implementing them in a lower-level programming language such as C++.

As it stands, the proposed expectation-based ZIP scan statistic could prove a useful tool for outbreak detection for problems and data for which it is believed suitable. Specifically, it may be of use in countries where reporting standards or medical facilities are not fully equipped to diagnose certain diseases, leading to underreporting or lack of reports for some regions. More generally, it could be included in the host of methods—scan statistics and others—that are employed in routine disease surveillance by public health authorities in most countries. Possible future extensions of this scan statistic could be an outbreak model that also affects the excess zero probability of the ZIP distribution—there exists a few ideas on this point, but they are likely to require an even larger computational effort. It could also be of interest to find other fields of application for the ZIP scan statistic, where perhaps excess zero counts have a natural explanation and are empirically evident.

Bibliography

Abrams, A. M., Kulldorff, M. & Kleinman, K. (2006), 'Empirical/asymptotic p-values for Monte Carlo-based hypothesis testing: an application to cluster detection using the scan statistic', *Advances in Disease Surveillance*.

Abrams, A. M., Kulldorff, M. & Kleinman, K. (2010), 'Gumbel based p-value approximations for spatial scan statistics', *International Journal of Health Geographics*.

Assuncao, R., Costa, M., Tavares, A. & Ferreira, S. (2006), 'Fast detection of arbitrarily shaped disease clusters', *Statistics in Medicine* 25(5), 723–742.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015*a*), 'Fitting linear mixed-effects models using `lme4`'. ArXiv e-print; in press, *Journal of Statistical Software*.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015*b*), `lme4`*: Linear mixed-effects models using Eigen and S4*. R package version 1.1-9.

Bivand, R. S., Pebesma, E. J. & Gomez-Rubio, V. (2013), *Applied spatial data analysis with R, Second edition*, Springer, NY.

Cançado, A. L., da Silva, C. Q. & da Silva, M. F. (2014), 'A spatial scan statistic for zero-inflated Poisson process', *Environmental and Ecological Statistics* 21(4), 627–650.

Carrel, M., Voss, P., Streatfield, P. K., Yunus, M. & Emch, M. (2010), 'Protection from annual flooding is correlated with increased cholera prevalence in Bangladesh: a zero-inflated regression analysis', *Environmental Health* 9(13), 13–21.

Centers for Disease Control and Prevention (CDC) (2015), 'Cryptosporidium', `http://www.cdc.gov/parasites/crypto/illness.html`.

Christiansen, L. E., Andersen, J. S., Wegener, H. C. & Madsen, H. (2006), 'Spatial scan statistics using elliptic windows', *Journal of Agricultural, Biological, and Environmental statistics* 11(4), 411–424.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society, Series B* 39(1), 1–38.

Dowle, M., Short, T., Lianoglou, S. & Srinivasan, A. (2014), `data.table`*: Extension of data.frame*. R package version 1.9.4.

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

Duczmal, L., Cancado, A. L. F., Takahashi, R. H. C. & Bessegato, L. F. (2007), 'A genetic algorithm for irregularly shaped spatial scan statistics', *Computational Statistics & Data Analysis* 52(1), 43–52.

Duczmal, L., Moreira, G. J., Burgarelli, D., Takahashi, R. H., Magalhães, F. C. & Bodevan, E. C. (2011), 'Voronoi distance based prospective space-time scans for point data sets: a dengue fever cluster analysis in a southeast Brazilian town', *International Journal of Health Geographics* 10(1), 29.

Dwass, M. (1957), 'Modified randomization tests for nonparametric hypotheses', *Annals of Mathematical Statistics* 28(1), 181–187.

Gertler, M., Dürr, M., Renner, P., Poppert, S., Askar, M., Breidenbach, J., Frank, C., Preußel, K., Schielke, A., Werber, D., Chalmers, R., Robinson, G., Feuerpfeil, I., Tannich, E., Gröger, C., Stark, K. & H, W. (2015), 'Outbreak of Cryptosporidium hominis following river flooding in the city of Halle (Saale), Germany, August 2013', *BMC Infectious Diseases*.

Glaz, J., Pozdnyakov, V. & Wallenstein, S. (2009), *Scan Statistics: Methods and Applications*, Birkhäuser Boston.

Hilbe, J. M. (2011), *Negative Binomial Regression*, 2 edn, Cambridge University Press.

Höhle, M. (2007), '`surveillance`: An R package for the monitoring of infectious diseases', *Computational Statistics* 22(4), 571–582.

Höhle, M., Meyer, S. & Paul, M. (2015), `surveillance`*: Temporal and Spatio-Temporal Modeling and Monitoring of Epidemic Phenomena*. R package version 1.9-1.

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

Kleinman, K., Abrams, A. M., Kulldorff, M. & Platt, R. (2005), 'A model-adjusted space-time scan statistic with an application to syndromic surveillance', *Epidemiology and Infection* 133(03), 409–419.

Kleinman, K., Lazarus, R. & Platt, R. (2004), 'A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism', *American Journal of Epidemiology* 159(3), 217–224.

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

Kulldorff, M. (1999), Spatial scan statistics: models, calculations, and applications, *in* 'Scan statistics and applications', Springer, pages 303–322.

Kulldorff, M. (2001), 'Prospective time periodic geographical disease surveillance using a scan statistic', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 164(1), 61–72.

Kulldorff, M. & Nagarwalla, N. (1995), 'Spatial disease clusters: Detection and inference', *Statistics in Medicine* 14(8), 799–810.

Kulldorff, M., Heffernan, R., Hartman, J., Assunçao, R. & Mostashari, F. (2005), 'A space-time permutation scan statistic for disease outbreak detection', *PLoS Medicine* 2(3), e59.

Kulldorff, M., Tango, T. & Park, P. J. (2003), 'Power comparisons for disease clustering tests', *Computational Statistics & Data Analysis* 42(4), 665–684.

Lambert, D. (1992), 'Zero-inflated poisson regression, with an application to defects in manufacturing', *Technometrics* 34(1), 1–14.

Lawless, J. F. (1987), 'Negative binomial and mixed poisson regression', *The Canadian Journal of Statistics* 15(3), 209–225.

Lehmann, E. L. & Romano, J. P. (2008), *Testing Statistical Hypotheses*, 3 edn, Springer.

Moni Bidin, C., de la Fuente Marcos, R., de la Fuente Marcos, C. & Carraro, G. (2010), 'Not an open cluster after all: the NGC 6863 asterism in Aquila', *Astronomy & Astrophysics*.

Naus, J. (1963), Clustering of Random Points in the Line and Plane, PhD thesis, Rutgers University.

Naus, J. (1965*a*), 'Clustering of random points in two dimensions', *Biometrika* 52, 263–267.

Naus, J. (1965*b*), 'The distribution of the size of the maximum cluster of points on a line', *Journal of the American Statistical Association* 60(310), 532–538.

Neill, D. B. (2006), Detection of spatial and spatio-temporal clusters, PhD thesis, Carnegie Mellon University.

Neill, D. B. (2009*a*), 'An empirical comparison of spatial scan statistics for outbreak', *International Journal of Health Geographics*.

Neill, D. B. (2009*b*), 'Expectation-based scan statistics for monitoring spatial time series data', *International Journal of Forecasting* 25(3), 498–517.

Neill, D. B. & Cooper, G. F. (2010), 'A multivariate Bayesian scan statistic for early event detection and characterization', *Machine Learning* 79(3), 261–282.

Neill, D. B. & Moore, A. W. (2006), Methods for detecting spatial and spatio-temporal clusters, *in* M. M. Wagner, A. W. Moore & R. M. Aryel, eds, 'Handbook of Biosurveillance', Academic Press.

Neill, D. B., Moore, A. W. & Cooper, G. F. (2006), 'A Bayesian spatial scan statistic', *Advances in Neural Information Processing Systems* 18, 1003.

Neill, D. B., Moore, A. W., Sabhnani, M. & Daniel, K. (2005), Detection of emerging space-time clusters, *in* 'Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining', ACM, pages 218–227.

Patil, G. & Taillie, C. (2004), 'Upper level set scan statistic for detecting arbitrarily shaped hotspots', *Environmental and Ecological statistics* 11(2), 183–197.

Pebesma, E. J. & Bivand, R. S. (2005), 'Classes and methods for spatial data in R', *R News* 5(2), 9–13.

Robert Koch Institute: SurvStat@RKI 2.0 (2015), 'Disease outbreaks', https://survstat.rki.de.

Salmon, M., Schumacher, D. & Höhle, M. (2015), 'Monitoring count time series in r: Aberration detection in public health surveillance', http://arxiv.org/abs/1411.1292. Accepted for Journal of Statistical Software.

Surti, S. & Karp, J. S. (2010), 'Application of a generalized scan statistic model to evaluate TOF PET images', *IEEE Transactions on Nuclear Science* 58(1), 99–104.

Takahashi, K., Kulldorff, M., Tango, T. & Yih, K. (2008), 'A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring', *International Journal of Health Geographics* 7(1), 14.

Tango, T. & Takahashi, K. (2005), 'A flexibly shaped spatial scan statistic for detecting clusters', *International Journal of Health Geographics* 4(1), 11.

Tango, T., Takahashi, K. & Kohriyama, K. (2011), 'A space-time scan statistic for detecting emerging outbreaks', *Biometrics* 67(1), 106–115.

Vergne, T., Paul, M. C., Chaengprachak, W., Durand, B., Gilbert, M., Dufour, B., Roger, F., Kasemsuwan, S. & Grosbois, V. (2014), 'Zero-inflated models for identifying disease risk factors when case detection is imperfect: Application to highly pathogenic avian influenza H5N1 in Thailand', *Preventive Veterinary Medicine* 114(1), 28–36.

Wickham, H. (2014), *Advanced R*, Chapman and Hall/CRC.

Wickham, H. (2015), *R Packages*, O'Reilly Media.

World Health Organization (2015), 'Disease outbreaks', http://www.who.int/topics/disease_outbreaks/en/.

A EM Algorithm for ZIP Parameters

Consider a sample of n observations $y_1, \ldots, y_n$ and known corresponding excess zero indicators $d_1, \ldots, d_n$. I.e. $d_i = 1$ if observation $y_i = 0$ is an excess zero, and $d_i = 0$ if y_i is a value generated from the Poisson component of the zero-inflated Poisson distribution. From Equation (3.2.6), we can see that the log-likelihood function for the parameters p and μ of the ZIP distribution with this data is given by

$$\ell(p, \mu | \{y_i\}, \{d_i\}) = \sum_{i=1}^{n} \left[d_i \log(p) + (1 - d_i) \log\left((1-p) e^{-\mu} \frac{\mu^{y_i}}{y_i!} \right) \right].$$

Taking partial derivatives w.r.t. p and setting the result equal to zero, we get

$$\frac{\partial}{\partial p} \ell(p, \mu | \{y_i\}, \{d_i\}) = \sum_{i=1}^{n} \left[\frac{d_i}{p} - \frac{1 - d_i}{1 - p} \right] = 0 \quad \Leftrightarrow \quad \hat{p} = \frac{1}{n} \sum_{i=1}^{n} d_i.$$

Similarly for μ, we get

$$\frac{\partial}{\partial \mu} \ell(p, \mu | \{y_i\}, \{d_i\}) = \sum_{i=1}^{n} \left[\frac{(1 - d_i) y_i}{\mu} - (1 - d_i) \right] = 0 \quad \Leftrightarrow \quad \hat{\mu} = \frac{\sum_{i=1}^{n} (1 - d_i) y_i}{\sum_{i=1}^{n} (1 - d_i)}.$$

Now for the EM algorithm, let each d_i be replaced by its unknown, random counterpart δ_i, and let $\hat{p}^{(k)}$ and $\hat{\mu}^{(k)}$ be the parameter estimates in the kth iteration of the algorithm. Define

$$\begin{aligned} Q(p, \mu | \hat{p}^{(k)}, \hat{\mu}^{(k)}) &\coloneqq \mathrm{E}\left[\ell(p, \mu | \{y_i\}, \{\delta_i\}) \middle| \{y_i\}; \hat{p}^{(k)}, \hat{\mu}^{(k)} \right] \quad \text{(expectation is w.r.t. all } \delta_i\text{)} \\ &= \sum_{i=1}^{n} \mathrm{E}\left[\delta_i \middle| \{y_i\}; \hat{p}^{(k)}, \hat{\mu}^{(k)} \right] \log(p) \\ &\quad + \sum_{i=1}^{n} (1 - \mathrm{E}\left[\delta_i \middle| \{y_i\}; \hat{p}^{(k)}, \hat{\mu}^{(k)} \right]) \log\left((1-p) e^{-\mu} \frac{\mu^{y_i}}{y_i!} \right). \end{aligned}$$

As in Section 3.2.3, we can see that

$$\begin{aligned}\mathrm{E}\left[\delta_i \middle| \{y_i\}; \hat{p}^{(k)}, \hat{\mu}^{(k)}\right] &= \mathrm{P}(\delta_i = 1 | Y_i = y_i; \hat{p}^{(k)}, \hat{\mu}^{(k)}) \\ &= \frac{\hat{p}^{(k)}}{\hat{p}^{(k)} + (1 - \hat{p}^{(k)}) \exp\left(-\hat{\mu}^{(k)}\right)} \mathbb{1}\{y_i = 0\} \\ &=: \hat{\delta}_i^{(k)},\end{aligned}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. Differentiating Q w.r.t. p and μ, and solving for these parameters, then leads to the same estimators as those we derived for the full likelihood, except that the values d_i are replaced by their estimates $\hat{\delta}_i^{(k)}$. The EM algorithm thus becomes

E-step at iteration k: set

$$\hat{\delta}_i^{(k)} = \frac{\hat{p}^{(k)}}{\hat{p}^{(k)} + (1 - \hat{p}^{(k)}) \exp\left(-\hat{\mu}^{(k)}\right)} \mathbb{1}\{y_i = 0\}. \tag{A.0.1}$$

M-step at iteration $k + 1$: set

$$\hat{p}^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{\delta}_i^{(k)} \tag{A.0.2}$$

$$\hat{\mu}^{(k+1)} = \frac{\sum_{i=1}^{n} (1 - \hat{\delta}_i^{(k)}) y_i}{\sum_{i=1}^{n} (1 - \hat{\delta}_i^{(k)})}. \tag{A.0.3}$$

With sensible initial values, these steps are repeated until a chosen convergence criterion has been met.

B Simulation Plots

This appendix contains plots of spatial precision, recall, and F-scores for simulated outbreaks with severities $q \in \{2, 2.5\}$, that were not included in Chapter 4.

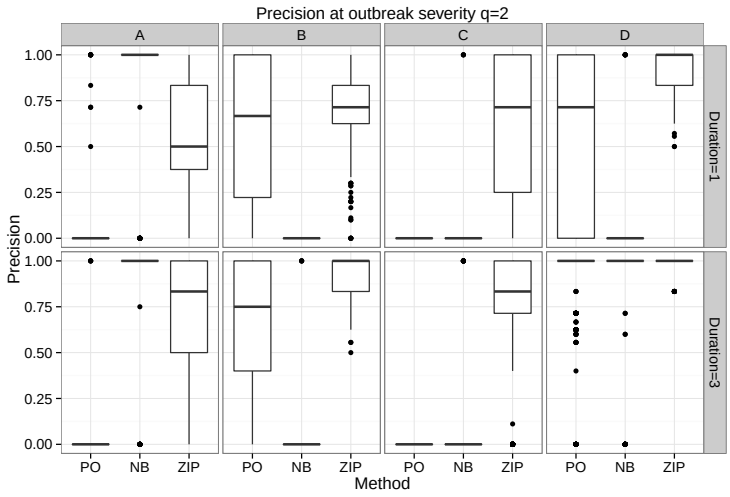

Figure B.1 Boxplot of the spatial recall calculated for the 1000 simulated outbreaks, for the scenarios in which the outbreak severity q is set at 2.

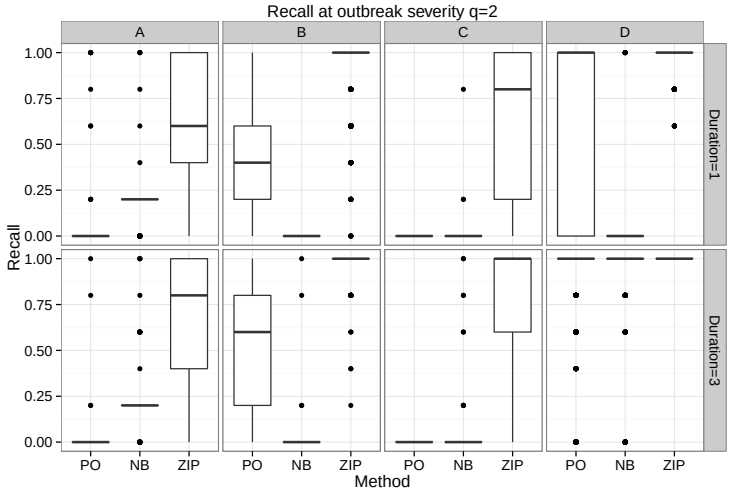

Figure B.2 Boxplot of the spatial recall calculated for the 1000 simulated outbreaks, for the scenarios in which the outbreak severity q is set at 2.

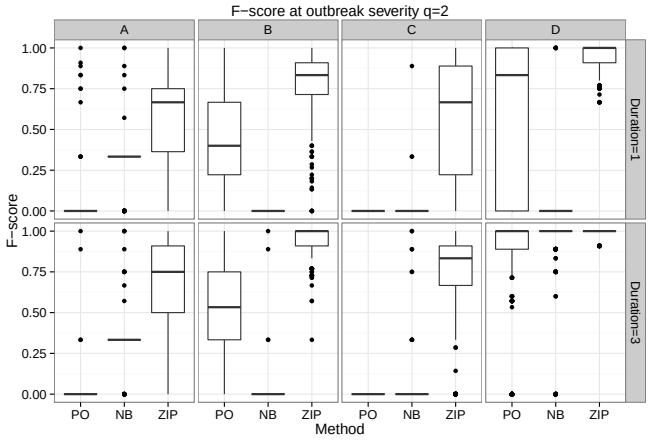

Figure B.3 Boxplot of the spatial recall calculated for the 1000 simulated outbreaks, for the scenarios in which the outbreak severity q is set at 2.

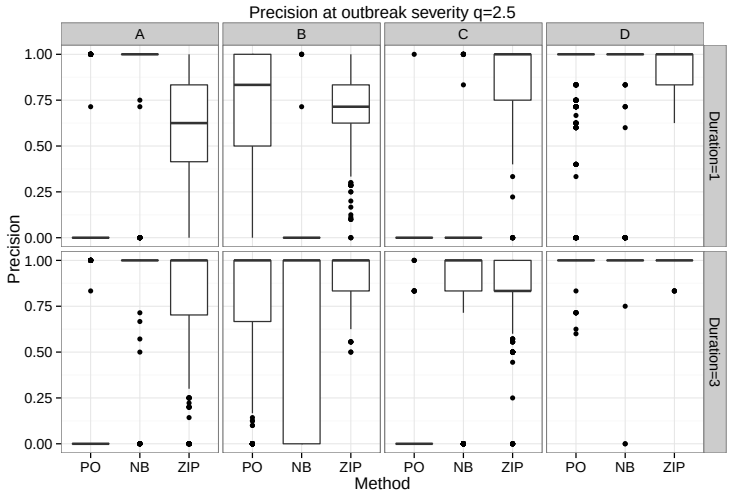

Figure B.4 Boxplot of the spatial recall calculated for the 1000 simulated outbreaks, for the scenarios in which the outbreak severity q is set at 2.5.

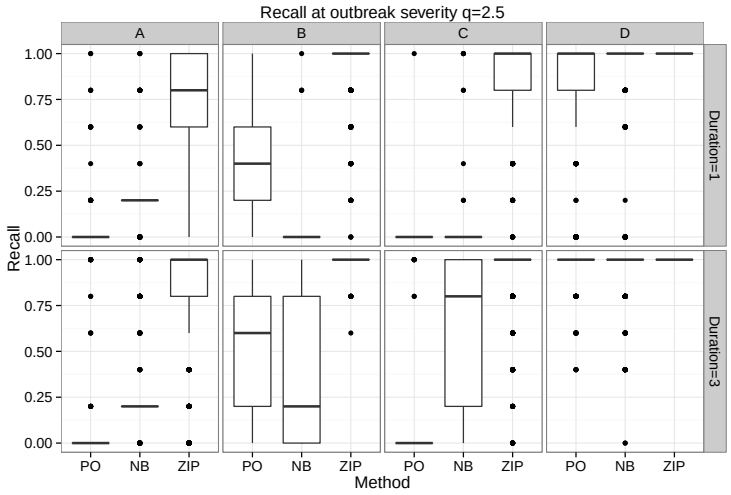

Figure B.5 Boxplot of the spatial recall calculated for the 1000 simulated outbreaks, for the scenarios in which the outbreak severity q is set at 2.5.

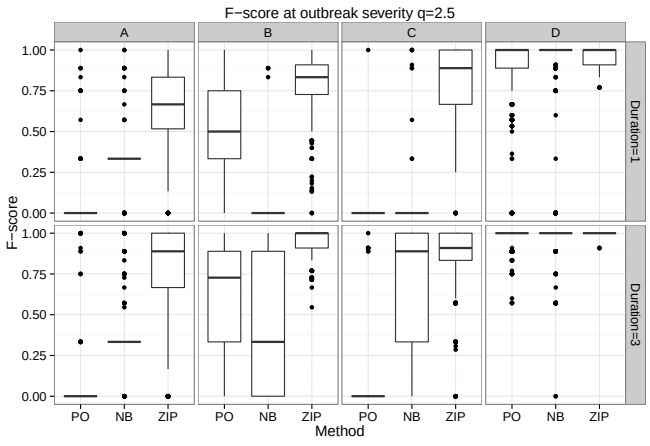

Figure B.6 Boxplot of the spatial recall calculated for the 1000 simulated outbreaks, for the scenarios in which the outbreak severity q is set at 2.5.

C An R Function to Fit a ZIP Mixed Model

In this chapter an R function for the fitting of a zero-inflated Poisson mixed model, as given in Chapter 5. It is important to note that this is only a tiny portion of the code that has been written for this thesis: the code for the R package `scanstatistics` that has been created for this thesis is available at https://github.com/BenjaK/scanstatistics. Further, over 1000 lines of code exist (and is available upon request) to produce the results in chapters 4 and 5.

The following function fits the ZIP mixed model:

```
#' Fit a simple mixed model for a ZIP distribution.
#'
#' Fit a simple mixed model for a zero-inflated Poisson distribution, for which
#' the log of the conditional expected value of the Poisson component is given
#' by \eqn{\mu_{it} = \gamma + a_i} and the logit of the excess zero probability
#' is given by \eqn{p_{it} = \rho + b_i}, where \eqn{i} is a group-level index.
#' The function was inspired by code for similar purposes found at:
#' https://groups.nceas.ucsb.edu/non-linear-modeling/projects/owls/R/
#' @param dt A \code{data.table} with columns \code{count} and \code{location},
#'    and a third column such as \code{time}.
#' @param maxitr An integers for the maximum number of iterations to perform.
#' @param tol Scalar; absolute convergence criterion.
#' @inheritParams lme4::glmer
#' @return A list with components
#' \describe{
#'   \item{poisson}{An object of class \code{glmerMod}; the fitted Poisson GLMM}
#'   \item{binary}{An object of class \code{glmerMod}; the fitted logistic GLMM}
#' }
fit_zip_glmm <- function(dt, maxitr = 20, tol = 1e-6, nAGQ = 1L, verbose = 0L) {
  # Get indices of zeros in data
  zero_idx <- dt[, .I[count == 0]]
  zero_ps <- numeric(dt[, .N])
  zero_ps[zero_idx] <- 1 / (1 + exp(-1)) # Initial value for excess zero probs

  # Separate data.table for binary data
  bin_data <- copy(dt)
  bin_data[count > 0, count := 1L]

  delta <- 1
  itr <- 1
  while (delta > tol & itr < maxitr) {
    old_zero_ps <- zero_ps

```

```r
    # Fit logistic GLMM for binary data
    bin_mod <- glmer(count ~ 1 + (1 | location),
                     data = bin_data, family = binomial, nAGQ = nAGQ)
    bin_fit <- fitted(bin_mod)

    # Poisson GLMM for count data, with weights given by 1 minus current
    # excess zero probability estimate
    count_mod <- glmer(count ~ 1 + (1 | location),
                       family = poisson(link = "log"), data = dt,
                       weights = 1 - zero_ps)
    count_fit <- fitted(count_mod)

    # Update excess zero probability estimates
    pz <- bin_fit[zero_idx]
    zero_ps[zero_idx] <- pz / (pz + (1 - pz) * exp(-count_fit[zero_idx]))

    delta <- max(abs(zero_ps - old_zero_ps))
    itr <- itr + 1
    if (verbose > 0) print(paste0("Iteration = ", itr - 1, ", delta = ", delta))
  }
  list(poisson = count_mod, binary = bin_mod)
}
```

R code