A BAYESIAN NETWORK MODEL FOR SPATIO-TEMPORAL EVENT SURVEILLANCE

by

Xia Jiang

B.S., Southern Institute of Metallurgy, 1989

M.S., Rose-Hulman Institute of Technology, 1997

Advisor: Gregory F. Cooper, Biomedical Informatics

Submitted to the Graduate Faculty of
the Department of Biomedical Informatics in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

This dissertation was presented

by

Xia Jiang

It was defended on

August 28, 2008

and approved by

Gregory F. Cooper, Associate Professor, Biomedical Informatics

Wendy Chapman, Assistant Professor, Biomedical Informatics

Milos Hauskrecht, Associate Professor, Computer Science

Daniel Neill, Assistant Professor, School of Public Policy and Mgmt., Carnegie Mellon

Dissertation Director: Gregory F. Cooper, Associate Professor, Biomedical Informatics

A BAYESIAN NETWORK MODEL FOR SPATIO-TEMPORAL EVENT SURVEILLANCE

Xia Jiang, PhD

University of Pittsburgh, 2008

Event surveillance involves analyzing a region in order to detect patterns that are indicative of some event of interest. An example is the monitoring of information about emergency department visits to detect a disease outbreak. Spatial event surveillance involves analyzing spatial patterns of evidence that are indicative of the event of interest. A special case of spatial event surveillance is spatial cluster detection, which searches for subregions in which the count of an event of interest is higher than expected. Temporal event surveillance involves monitoring for emerging temporal patterns. Spatio-temporal event surveillance involves joint spatial and temporal monitoring.

When the events observed are of direct interest, then analyzing counts of those events is generally the preferred approach. However, in event surveillance we often only observe events that are indirectly related to the events of interest. For example, during an influenza outbreak, we may only have information about the chief complaints of patients who visited emergency departments. In this situation, a better surveillance approach may be to model the relationships among the events of interest and those observed.

I developed a high-level Bayesian network architecture that represents a class of spatial event surveillance models, which I call BayesNet-S. I also developed an architecture that represents a class of temporal event surveillance models called BayesNet-T. These Bayesian network architectures are combined into a single architecture that represents a class of spatio-temporal models called BayesNet-ST. Using these architectures, it is often possible to construct a temporal, spatial, or spatio-temporal model from an existing Bayesian network

event-surveillance model that is non-spatial and non-temporal. *My general hypothesis is that when an existing model is extended to incorporate space and time, event surveillance will be improved.*

PANDA-CDCA (PC) (Cooper et al., 2007) is a non-temporal, non-spatial disease outbreak detection system. I extended PC both spatially and temporally. *My specific hypothesis is that each of the spatial and temporal extensions of PC will perform outbreak detection better than does PC, and that the combined use of the spatial and temporal extensions will perform better than either extension alone.*

The experimental results obtained in this research support this hypothesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

A new Bayesian network model for spatio-temporal event surveillance is developed and evaluated in this thesis. This chapter first describes spatio-temporal event surveillance in general, and it then summarizes the contributions of this thesis.

1.1 EVENT SURVEILLANCE

Event surveillance consists of analyzing a region in order to detect patterns that are indicative of some event of interest. As examples, we may look for patterns that are indicative of a forthcoming disaster or a disaster that is in its early stages. Examples of such disasters include hurricanes, terrorist attacks, and outbreaks of diseases. A classic example of event surveillance involves monitoring some geographical region in order to detect a disease outbreak. In this thesis, the focus will be on disease outbreaks.

(Le Strat and Carrat, 1999) define an **epidemic** as "the occurrence of a number of cases of a disease, in a given period of time in a given population, that exceeds the expected number," while (Last, 2000) defines a **disease outbreak** as "an epidemic limited to a localized increase, as for example in a village, town, or institution." On a given day, the number of cases could by chance exceed the expected number, and then return to normalcy. Ordinarily, this would not be considered a disease outbreak. A disease outbreak is characterized by an increasing trend (with daily fluctuations) in cases until some peak is reached, then a decline, and then possibly an increase to a second peak, and so on. **Disease outbreak surveillance**, also called **disease outbreak detection** and **biosurveillance**, consists of monitoring a community in order to recognize early the onset of a disease outbreak. See

(Buckeridge, 2007), (Buckeridge et al., 2005a), (Bravata et al., 2004), and (Feinberg and Shmueli, 2005) for a review of biosurveillance.

1.1.1 Temporal Event Surveillance

Our pattern of interest may or may not be emerging in time. For example, we may be interested in whether there is a cluster of a particular type of tree in a forest. In this case there is no change in the pattern from one time period to the next (at least in the units of time we are considering). On the other hand, in applications such as disease outbreak detection, the pattern is emerging in time. A method for **non-temporal event surveillance** only looks at data from the most recent time period. Such a method can be used to investigate an emerging pattern such as a disease outbreak. However, the analysis would not look at data from previous time periods. A method for **temporal event surveillance** looks for **emerging patterns** by analyzing how the situation has changed recently in time. The analysis is based not only on the data from the most recent time period, but also on data from previous time periods.

1.1.2 Spatial Event Surveillance

In non-spatial event surveillance, an entire region is monitored globally. For example, if we were monitoring whether a disease outbreak was occurring in a particular county, we would monitor the entire county globally, without considering the possibility of localized outbreaks in subregions.

If an outbreak was occurring in a small subregion of a county and the entire county was monitored globally, the outbreak may go undetected until it spread to a larger subregion. In **spatial event surveillance**, we search for patterns in spatial subregions. That is, we individually monitor both small and large subregions of the region of interest. In this way, we not only may detect an emerging event sooner, but we may also learn its location. For example, Figure 1.1 shows how a disease outbreak might emerge in Allegheny County, Pennsylvania. If we were investigating a spatial subregion near the center of the county, we might detect the outbreak earlier and pinpoint its location.

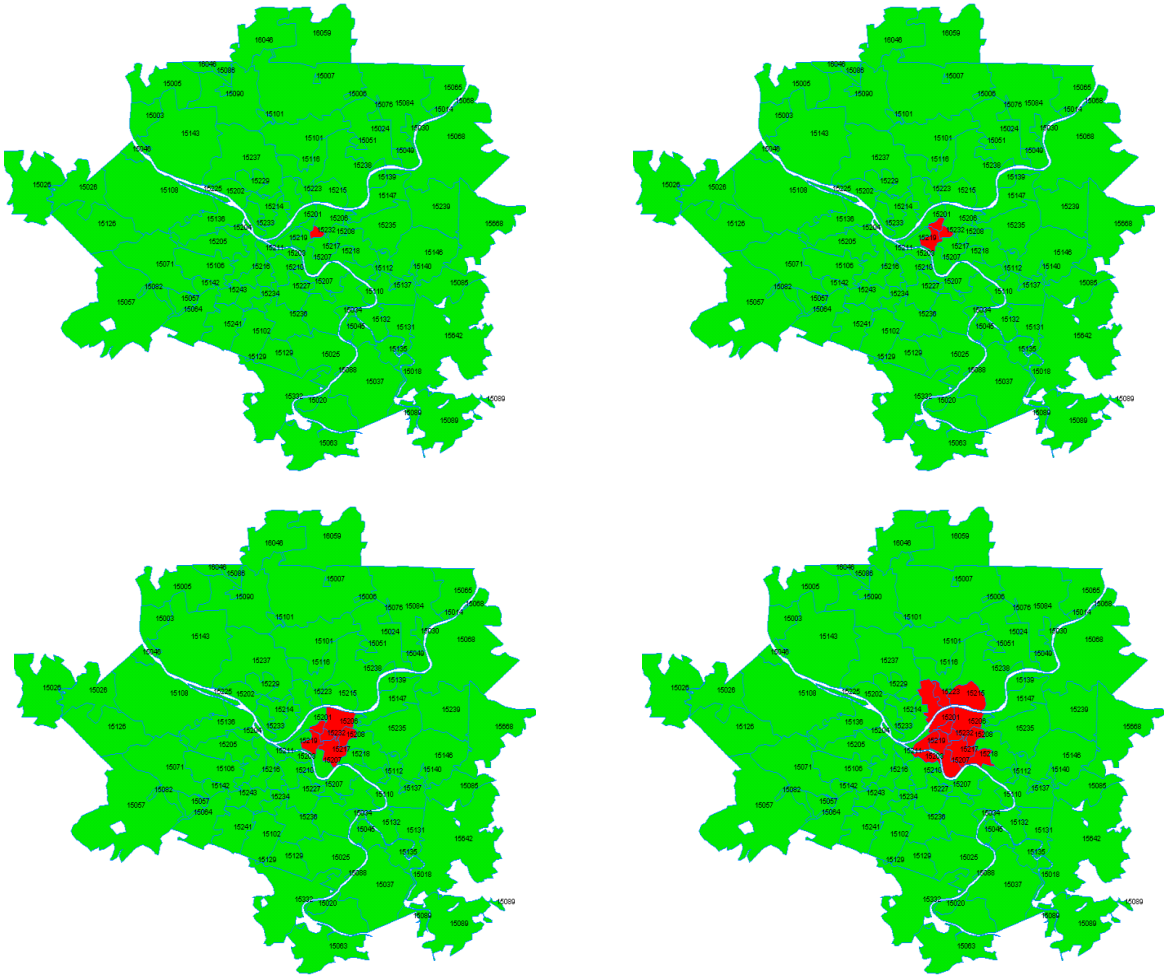


Figure 1.1: The progression of a fictitious outbreak in Allegheny County. The progression is from top left to top right to bottom left to bottom right.

1.1.3 Spatio-Temporal Event Surveillance

In **spatio-temporal event surveillance**, we look for patterns that are emerging in time both by investigating subregions of a region of interest and by analyzing how the situation is changing in time.

1.1.4 Importance of Spatio-Temporal Event Surveillance

Early, reliable, and accurate detection of disease outbreaks remains an important research topic. Even modest improvements in disease outbreak detection could have significant impact on public health in terms of lives saved and reduced economic cost. The induced long-term economic costs were estimated to be as high as 250 million dollars per hour for some types of outbreaks (Kaufmann et al., 1997; Wagner et al., 2001). We need to further improve our outbreak detection systems to detect outbreaks earlier and more reliably than currently done. We also need to further improve the capability of our detection systems in pinpointing the geographical subregion where an outbreak is taking place. Ideally, we want a system that is not only able to pinpoint the geographical subregion of an outbreak but also able to identify the type/bioagent of the outbreak among all possible outbreaks.

1.1.5 Spatial Cluster Detection and Spatial Event Surveillance

Spatial cluster detection is one statistical technique used for spatial event surveillance. Methods for **spatial cluster detection** attempt to locate spatial subregions of some larger region where the count of occurrences of some event is higher in one subregion relative to other subregions. The classic technique for analyzing these counts is the spatial scan statistic (Kulldorff, 1997, 1999). In the case of disease outbreak detection, we want to find clusters of disease cases so as to pinpoint where the outbreak is occurring. Other applications of spatial cluster detection include mining astronomical data, medical imaging, and military surveillance. In all these applications, the goal is to identify the location, shape, and size of possible clusters, and to determine how likely it is that the cluster is due to the event with which we are concerned (e.g., a disease outbreak) verses how likely it is that the cluster is

merely a chance occurrence.

1.1.6 Bayesian Networks and Spatial Event Surveillance

When the events observed are the events of interest, then directly analyzing counts is likely to be a good surveillance method. However, in event surveillance in general, and disease-outbreak detection in particular, we often may only observe events that are *related* to the events of interest. As an example, when we are interested in whether there is an outbreak of a certain disease, we observe individuals with symptoms of the disease rather than more direct evidence of the disease being present. Instead of using a summary statistic, we may obtain better results if we model the relationships among the event of interest, and the observable events using a Bayesian network (see Section 2.1 for an introduction to Bayesian networks). The Bayesian network is then used to determine the posterior probability of each subregion containing the event of interest. A strength of this method is that it can model multiple causes of the observed occurrences. For example, in disease outbreak detection, it can model any number of possible disease outbreaks using a single Bayesian network. Such a Bayesian network approach is developed in this thesis.

1.2 CONTRIBUTIONS OF THIS THESIS

A high-level Bayesian network architecture, representing a class of spatial event surveillance models called BayesNet-S, is developed. Then a high-level Bayesian network architecture representing a class of temporal event surveillance models called BayesNet-T is developed. These Bayesian network architectures are then combined into one high-level Bayesian network architecture that represents a class of spatio-temporal model called BayesNet-ST. Using these high-level Bayesian network architectures, it is often possible to construct a temporal, spatial, or spatio-temporal model from an existing Bayesian network model for non-spatial, non-temporal event surveillance. This is called extending the existing model. *My general hypothesis is that event surveillance will be improved when an existing model is extended to*

incorporate space and time within domains where space and time are relevant.

PANDA-CDCA (PC) (Cooper et al., 2007) is a previously developed disease-outbreak-detection system that uses a Bayesian network to model the relationships among the events of interest and those observed. However, PC does not use a temporal or spatial model of disease outbreaks. Using my high-level Bayesian network architectures, I have extended the PC algorithmic framework to include spatial and temporal models, and I have extended the PC disease-outbreak domain model to consider spatial and temporal aspects of disease outbreaks. *My specific hypothesis is that both the spatial and temporal extensions of PC will perform outbreak detection better than does PC and that the combined use of the spatial and temporal extensions will perform better than either extension alone.* Chapter 4 describes experiments that evaluate this hypothesis.

1.2.1 Contribution to Spatial Event Surveillance

A BayesNet-S model may show better detection performance than the spatial scan statistic (Kulldorff, 1997) because such a model has the following potential advantages. A BayesNet-S model can readily include multinomial variables, whereas the spatial scan statistic cannot. The spatial scan statistic only investigates whether a cluster is occurring in a given subregion, whereas a BayesNet-S model can use a Bayesian network to model the causal mechanisms by which the clusters might occur. A BayesNet-S model can report the posterior probability of an outbreak in each subregion. Also, as will be shown in Section 2.3, the spatial scan statistic uses computationally expensive randomization methods to determine the likelihood of an outbreak.

1.2.2 Contribution to Temporal Event Surveillance

A system that looks only at each day's data might signal an outbreak one day and not signal it the next. Such a system will confuse the user as to whether or not there truly is an outbreak. For example, (Cooper et al., 2007) obtained such results when evaluating the ability of PC to detect a laboratory validated outbreak of influenza in Allegheny County. Under a false alarm rate of zero, PC detected influenza approximately one day before the first positive

viral cultures of influenza were taken. However, during the start of the influenza outbreak, the posterior probability of influenza fluctuated between very high and very low values. It seems that these fluctuations may be due to PC only considering the most recent 24 hours of data in performing outbreak detection. By extending PC to a BayesNet-T model, which is done in Section 3.3.2, the fluctuations of the posterior for influenza should be attenuated.

1.2.3 Contribution to Spatio-Temporal Event Surveillance

To my knowledge, BayesNet-ST is the first Bayesian network, spatio-temporal event surveillance model that has been developed. This model combines the strengths of BayesNet-S and BayesNet-T.

1.2.4 Contribution to Overall to the Field of Biomedical Informatics

Applications of my architectures extend beyond biosurveillance. My spatial architecture is applicable to many types of anomaly detection including medical imaging for the purpose of pathology detection. My temporal architecture is applicable to any type of monitoring that concerns a system which changes over time. For example, it may be used in a medical expert system, which is deployed in an intensive care unit (ICU), and which monitors changes in a patient's condition over time.

1.3 OUTLINE OF THIS THESIS

Chapter 2 provides background on statistical methods, Bayesian networks, and existing methods for event surveillance. The BayesNet-S, BayesNet-T, and BayesNet-ST classes of models are developed in Chapter 3. Furthermore, PC (Cooper et al., 2007) is extended to spatial, temporal, and spatio-temporal models in that chapter. Chapter 4 shows the results of experiments evaluating the performance of these extensions. The purpose of these experiments is to validate the hypotheses of this thesis. Chapter 5 further evaluates the performance of the extensions of PC by comparing their performance to that of state-of-

the-art event surveillance systems. In particular, the spatio-temporal extension of PC is compared to the spatio-temporal spatial scan statistic (Kulldorff, 2001; Kulldorff et al., 2005), which is discussed in Section 2.3.3.2. Finally, Chapter 6 offers some conclusions and suggestions for future research.

2.0 BACKGROUND

Methods that use frequentist and Bayesian statistics are presented and compared in this thesis. This chapter first discusses differences between frequentist and Bayesian methods, it then reviews Bayesian networks, and finally it provides an overview of methods that have previously been developed for event surveillance, with a focus on disease outbreak detection (biosurveillance).

2.1 FREQUENTIST VERSES BAYESIAN STATISTICS

A classic example of probability concerns tossing a coin. If the coin is symmetrical, we can use the Principle of Indifference to assign

$$P(\mathsf{Heads}) = P(\mathsf{Tails}) = 0.5$$

The **Principle of Indifference** says that outcomes are to be considered equiprobable if there is no reason to expect one over the other. Suppose that we toss a thumbtack. It can also land one of two ways. Because the thumbtack is not symmetrical, we have no reason to apply the Principle of Indifference and assign probabilities of 0.5 to both outcomes. In the case of the coin, when we assign $P(heads) = 0.5$, we are implicitly assuming that if we tossed the coin a large number of times it would land heads about half the time. (von Mises, 1919) used the limit of the fraction of heads as the definition of probability. That is, if n is the number of tosses and S_n is the number of times the thumbtack lands heads (on its flat end), then

$$P(\mathsf{Heads}) \equiv \lim_{n\to\infty} \frac{S_n}{n}.$$

This approach to probability is called the **relative frequency approach to probability**, and probabilities obtained using this approach are called **relative frequencies**. A **frequentist** is someone who feels this is the only way we can conceptualize probabilities. Note that, according to this approach, we can never know a probability for certain. For example, if we tossed a coin 10,000 times and it landed heads 7000 times, we could estimate that the probability is 0.7, but we could not know that this is the probability.

A strict frequentist approach to statistics does not manipulate probabilities because it assumes probabilities cannot be known for certain. Rather it infers confidence about unknown probabilities using techniques such as confidence intervals and hypothesis testing. A simple example follows. The technique illustrated in the example is covered in any elementary frequentist statistics text such as (Anderson, 2005).

Example 2.1. *Suppose a nut distributor says that on the average it puts* 3 *pounds of nuts in its* 3 *pound nut containers. We decide to investigate the claim by obtaining a random sample of size* $n = 40$ *nut containers. In this sample, we find that the average weight of the nuts is* $\bar{x} = 2.92$ *pounds, Suppose further that from years of previous data, we know that the standard deviation is* $\sigma = 0.3$*. We are interested in investigating whether the true mean* μ *is* ≥ 3 *because if this is the case the distributor is delivering on the average at least the amount it says. We call this event the null hypothesis and denote it* H_0*. We call the event that* μ *is* < 3 *the alternative hypothesis* H_A*. If the alternative hypothesis is true, the manufacturer is not supplying enough nuts. Formally, we have the following.*

$$\begin{aligned} H_0 &: \mu \geq 3 \\ H_A &: \mu < 3. \end{aligned}$$

If we assume a normal distribution and perform a Z-test, we obtain a p-value equal to 0.046*. This means that if* H_0 *is true, the probability of getting an average weight of* 2.92 *pounds or less is* 0.046*. Traditionally, such a p-value is interpreted as moderately strong evidence for rejecting the null hypothesis and thereby accepting the alternative hypothesis* H_A*.*

Notice in the previous example that probabilities were never known or manipulated.

If we tossed a thumbtack 10,000 times and it landed heads 7000 times, a frequentist could only obtained a confidence interval for the unknown probability of heads. However, in the **subjective approach to probability**, we can say that our belief concerning the outcome of heads on the next toss is exactly equal to 0.7, and this belief is our probability of heads. Often this belief is assessed by considering a fair gamble. That is, we would consider it fair to win \$0.30 if the thumbtack landed heads on the next toss and to lose $\$1 - \$0.30 = \$0.70$ if the thumbtack landed tails.[1] A **subjectivist** is someone who feels he or she can assign probabilities that represent his or her beliefs.

Since a subjectivist can "know" probabilities, the subjectivist can also manipulate them. This is ordinarily done using Bayes' Theorem, and so subjectivists are also called **Bayesians**. A Bayesian approach to statistics is one that infers unknown probabilities from known ones using Bayes' Theorem. The following is an example of Bayesian statistical inference.

Example 2.2. *Suppose a man takes the blood test ELISA (enzyme linked immunosorbent assay) which tests for the presence of HIV (human immunodeficiency virus), because the man is applying for a marriage license and the state requires this test. Suppose further that the true positive rate for the test is* 0.999*, and the false positive rate is* 0.002*. Our belief concerning the probabilities of the man testing positive, given the man either is or is not infected with HIV, are then as follows:*

$$P(ELISA = pos|HIV = yes) = 0.999$$

$$P(ELISA = pos|HIV = no) = 0.002.$$

Suppose further that 1 *in* 100,000 *men, who apply for a marriage license in this state, are infected with HIV. Our prior belief concerning the probability of the man being infected with HIV is then as follows:*

$$P(HIV = yes) = 0.00001.$$

Using Bayes' Theorem, we can now compute the posterior probability of the man being infected:

[1] For simplicity this example assumes the preference for money of this small amount is linear in the amount of money. If not, a simple modification will preserve the basic principle given in this example.

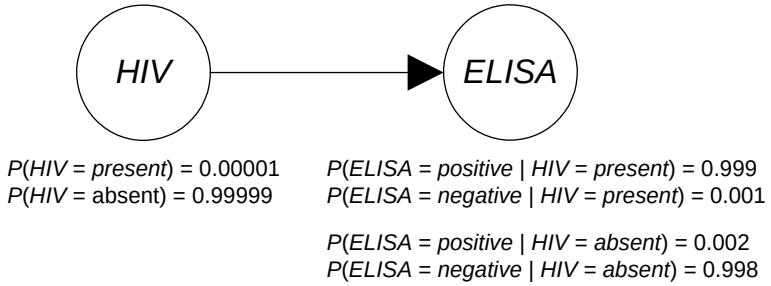


Figure 2.1: A two-node Bayesian network.

$$\begin{aligned} P(HIV = yes|ELISA = pos) &= \frac{P(pos|yes)P(yes)}{P(pos|yes)P(yes) + P(pos|no)P(no)} \\ &= \frac{(0.999)(0.00001)}{(0.999)(0.00001) + (0.002)(0.99999)} \\ &= 0.00497. \end{aligned}$$

See (Berry, 1996) for more on Bayesian statistics.

2.2 BAYESIAN NETWORKS

Bayesian networks will now be reviewed. See (Castillo et al., 2007; Kjaerulff and Madsen, 2008; Jensen, 1997; Jensen and Nielsen, 2007; Neapolitan, 1990, 2004; Pearl, 1988) for a detailed introduction to Bayesian networks.

In Example 2.2 we computed the probability of a man being infected with HIV given that he tested positive for HIV using Bayes' Theorem. We can represent the probabilities used in this computation in Figure 2.1, which is a two-node Bayesian network. In that figure, the random variables HIV and $ELISA$ are represented by nodes[2] in a directed acyclic graph, and the relationship between these variables is represented by an edge from HIV to $ELISA$.

[2]Nodes in a Bayesian network represents domain variables and in this dissertation I will use these terms interexchangeably.

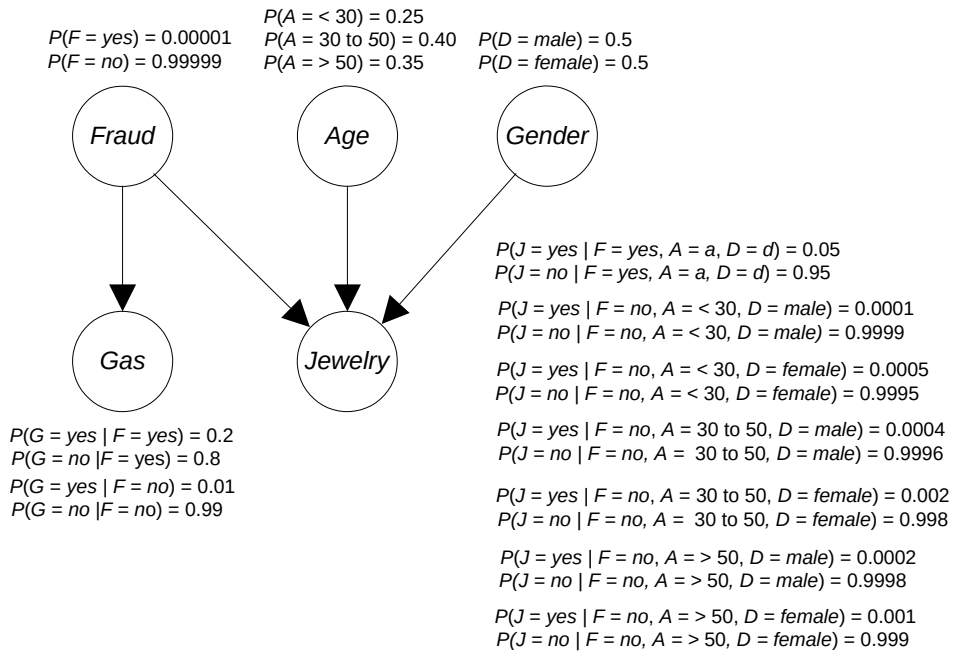


Figure 2.2: A Bayesian network for detecting credit card fraud.

Before proceeding, let us review some graph theory. Recall that a directed graph is a graph in which the edges have direction (an arrowhead), and a directed acyclic graph (DAG) is a directed graph in which there is no path from a node to itself. In a directed graph, node Y is called a parent of X if there is an edge from node Y to node X, and node Y is called a nondescendent of node X if there is no path from node X to node Y. For example, in Figure 2.2, *Gender* is a parent of Jewelry because there is an edge from *Gender* to $Jewelry$, while *Gas* is a nondescendent of *Gender* because there is no path from *Gender* to *Gas*.

In general, a Bayesian network consists of a DAG, whose edges represent relationships among random variables that are often causal; the prior probability distribution of every root variable in the DAG; and the conditional probability distribution of every non-root variable given each set of values of its parents. Figure 2.2 shows a more complex Bayesian network representing the causal relationships among variables related to credit card fraud (taken from (Heckerman, 1997)). Using this Bayesian network, we can determine conditional

probabilities of interest using the Bayesian network and a Bayesian network inference algorithm. For example, if a given individual is a male, is less than 30 years old, and jewelry was purchased using the individual's credit card, we can determine the conditional probability of the individual's credit card being used fraudulently. These inference algorithms exploit Bayes' Theorem and are efficient for a large class of Bayesian networks (Castillo et al., 2007; Kjaerulff and Madsen, 2008; Jensen, 1997; Jensen and Nielsen, 2007; Neapolitan, 1990, 2004; Pearl, 1988).

In a Bayesian network the product of the conditional probability distributions in the DAG must equal the joint probability distribution of the random variables. A formal definition of a Bayesian network and a theorem concerning this matter follow.

Definition 2.1. *Suppose we have a joint probability distribution P of the random variables in some set V and a DAG $\mathbb{G} = (\mathsf{V}, \mathsf{E})$. We say that $(\mathbb{G}, P)$ satisfies the* ***Markov condition*** *if for each variable $X \in \mathsf{V}$, X is conditionally independent of the set of all its nondescendents given the set of all its parents. That is, if the sets of parents and nondescendents of X are denoted by PA and ND, respectively, then for all values of X, ND, and PA*

$$P(X|\mathsf{ND}, \mathsf{PA}) = P(X|\mathsf{PA}).$$

If $(\mathbb{G}, P)$ satisfies the Markov condition, we call $(\mathbb{G}, P)$ a ***Bayesian network****.*

The proof of the following theorem can be found in (Neapolitan, 2004).

Theorem 2.2. *$(\mathbb{G}, P)$ satisfies the Markov condition (and therefore is a Bayesian network) if and only if P is equal to the product of its conditional distributions of all nodes given their parents in G, whenever these conditional distributions exist. That is, if our variables are $X_1, X_2, \ldots, X_n$, and PA_i is the set of parents of X_i, then*

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\mathsf{PA}_i).$$

Due to the previous theorem, Bayesian networks are often developed by first defining a DAG that satisfies the Markov condition relative to our belief about the probability distribution of the nodes in the DAG, and then determining the conditional probability distributions for this DAG. Often the DAG is a causal DAG, which is a DAG in which there is an edge from X to Y if and only if X is a direct cause of Y relative to the other nodes in the DAG. See (Neapolitan, 2004) for discussions as to why a causal DAG should often satisfy the Markov condition with the probability distribution of the variables in the DAG. The Bayesian networks in Figures 2.1 and 2.2 were developed by identifying causal edges. In general, however, the edges need not be causal. Rather it is only necessary that the Markov condition be satisfied.

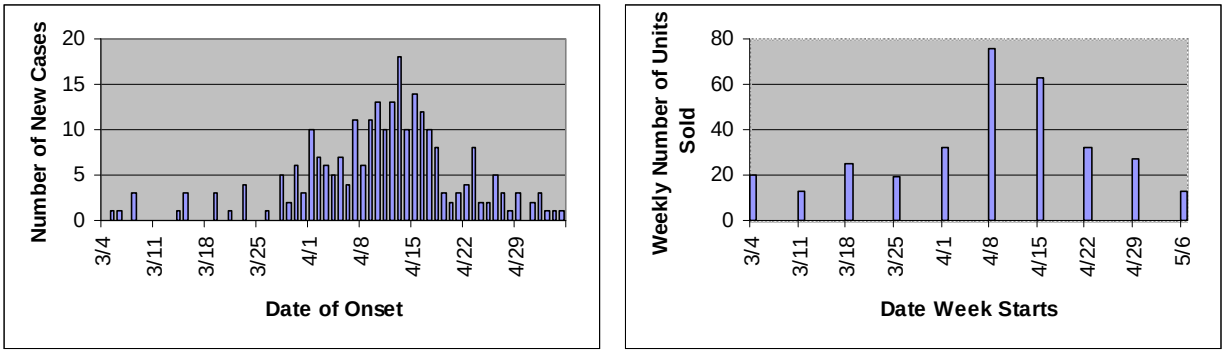
Methods for learning Bayesian networks from data have been developed (see (Neapolitan, 2004)).

2.3 OVERVIEW OF METHODS FOR EVENT SURVEILLANCE

This section provides an overview of methods that have previously been developed for event surveillance, with a focus on disease outbreak detection (biosurveillance). The purpose is not to be exhaustive, but rather to provide a representative overview of work in the field.

2.3.1 Methods for Non-Spatial Event Surveillance

2.3.1.1 Methods that Analyze Counts Often the count of occurrences of some phenomenon increases during a disease outbreak. For example, Figure 2.3 (a) shows an epidemic curve constructed from a sample of the population affected by a *Cryptosporidium* outbreak in North Battleford, Saskatchewan in spring, 2001. The outbreak was caused by a contamination of public drinking water. *Cryptosporidium* infection causes diarrhea. Figure 2.3 (b) shows the weekly counts of units of over-the-counter (OTC) antidiarrheal medicine sold at one pharmacy in North Battleford during the time period affected by the outbreak. The correlation between these two curves suggests that by monitoring OTC sales of such medicine



(a) (b)

Figure 2.3: An epidemic curve for a *Cryptosporidium* disease outbreak in North Battleford, Saskatchewan is in (a), while weekly OTC sales of antidiarrheal drugs at one pharmacy in North Battleford is in (b). The data for these curves ware obtained from (Stirling et al., 2001).

we can possibly detect a *Cryptosporidium* outbreak at an early stage. Similarly, the number of patients visiting the emergency department (ED) with respiratory symptoms ordinarily increases during an influenza[3] outbreak.

To monitor and analyze the counts, we first articulate a unit of time, which is ordinarily one day, but could be one hour, or any other unit. For the sake of discussion, in what follows it is assumed that the time unit is one day. A count is then obtained separately for each day.

Non-Temporal Methods:

Non-temporal methods look at counts from some recent period of time only, such as the previous 24 hours. One method for analyzing these daily counts is to first derive the mean μ and standard deviation σ of the daily counts over a period of time when no outbreak is presumed to be occurring, and fix these values in the outbreak detection system. An alert is then issued whenever the daily count exceeds μ by $k\sigma$, where k is usually 2 or 3. (Wong

[3]In the figures of this thesis we will use "flu" as a short name for "influenza".

and Moore, 2006) discuss problems with this method and improvements to it.

Temporal Methods:

Temporal methods detect an outbreak based on how the situation has changed recently in time. The determination of an outbreak is based not only on the count from the most recent day, but also on counts from previous days.

There are a number of temporal (time series) methods that look at the count of occurrences of a single phenomenon. Some of these methods are discussed in (Wong and Moore, 2006). (Moore et al., 2003) provide a summary of many known methods. Briefly, they include the Serfling method (Serfling, 1963; Tsui et al., 2001), the ARMA, ARIMA, and SARIMA models (Box et al., 1994; Hamilton, 1994), univariate hidden Markov models (Rabiner, 1989; Moore, 2001b), Kalman filters (Hamilton, 1994), support vector machines (Burgess, 1998; Moore, 2001c), and CUSUM (Bos and Fetherston, 1992). Other frequentist methods appear in (Burkom et al., 2007), (Buckeridge et al., 2005b), (Ries and Mandl, 2003), (Ries at al., 2003), and (Soneson and Bock, 2003). A Bayesian method is developed in (Jiang and Wallstrom, 2006).

CUSUM is one of the most widely used temporal methods, and it is used in an experiment in Section 5.3.2. Therefore, I provide some details of the CUSUM algorithm here. CUSUM analyzes the counts from the previous i time slots. Let μ_0 be the mean of the counts during some background period when no outbreak is occurring, σ be the standard deviation of the counts during the background period, and

$$H = d\sigma.$$

It is recommended that $d = 5$. Let $X_1, X_2, \ldots X_i$ be the counts from the past i time periods, and let

$$\begin{aligned} S_1 &= X_1 - \mu_0 \\ S_2 &= (X_1 - \mu_0) + (X_2 - \mu_0) = S_1 + (X_2 - \mu_0) \\ &\vdots \\ S_i &= \sum_{j=1}^{i} (X_j - \mu_0) = S_{i-1} + (X_i - \mu_0)\,. \end{aligned}$$

To determine when to signal an alarm, we look at this sequence:

$$\begin{aligned} C_1 &= \max(0, X_1 - (\mu_0 + K)) \\ &\vdots \\ C_i &= \max(0, C_{i-1} + X_i - (\mu_0 + K)). \end{aligned}$$

We signal an alarm whenever

$$C_i > H.$$

Temporal methods that look at several counts, each recording occurrences of a different phenomenon, appear in (Burkhom et al., 2004), (Burkhom et al., 2005), (Moore et al., 2006), and (Shmuel and Feinberg, 2006). (Reis et al., 2007) developed an epidemiological network model that monitors the relationships among different data streams instead of monitoring the data streams themselves. The Bayesian method developed in (Jiang and Wallstrom, 2006) can look at several counts, but in the implementation which they evaluated it did not. Methods that look at several counts are called **multivariate**.

2.3.1.2 Entity-Based Methods Rather than analyzing data aggregated over the entire population (i.e., daily counts of some observable events), another approach is to model the relationships among disease outbreaks and probabilistic properties of each individual in a population. This is an entity-based approach. By modeling each individual in the population, we can base our analysis on more information than that contained in a summary statistic such as the number of patients who visited the ED with respiratory symptoms on a given day. Methods that use this approach will now be presented.

Non-Temporal Methods:

PANDA-CDCA (PC) (Cooper et al., 2007) is a non-temporal method that models the CDC Category A diseases, namely, anthrax, plague, smallpox, tularemia, botulism and hemorrhagic fever, and also several diseases that may be confused with them, namely influenza, *Cryptosporidium*, and hepatitis (see http://www.bt. cdc.gov/agent/ agentlist-category.asp). PC consists of a large Bayesian network that contains a set of nodes for each individual in a

region. PC takes as input a time series of chief complaints, one for each ED patient in the region. There are 54 chief complaints, including a catchall category of "other". Each hour, based on the previous 24 hours (one day) of data, it outputs the posterior probability of each disease. PC not only can inform us if an outbreak is likely, but also what type of outbreak it might be. Note that, even though PC is run each hour and the data is collected over 24 hours, we still do not consider it a temporal model. The frequency with which we run the model does not determine whether we call it temporal. In order to label a surveillance method as temporal, we require that it consider patterns of evidence over time, not just evidence during a single time period. The details of PC are provided in Section 3.1.3.

BARD (Bayesian Aerosol Release Detector) (Hogan et al., 2007) is a Bayesian network, entity-based system designed to compute the posterior probability of an outdoor, wind-borne release of anthrax spores. BARD's goal is to perform earlier, more sensitive detection of wind-borne outbreaks by recognizing a characteristic dispersion pattern. It not only detects an outbreak, but characterizes it as wind-borne. Furthermore, it determines estimates of release location, quantity, and time.

Temporal Methods:

A predecessor to PC, PANDA (Population-wide ANomaly Detection and Assessment) (Cooper et al., 2004) is an entity-based, Bayesian network method that has a simple temporal and spatial model of an outbreak disease. PANDA is designed specifically to detect non-contagious outbreak diseases such as airborne anthrax or West Nile encephalitis. PANDA currently is able to detect disease outbreaks due to inhalational anthrax, and the only clinical evidence considered by this system is whether an individual presented to the ED with respiratory symptoms or not. The Bayesian network in PANDA contains a set of nodes for each individual in a region. These person nodes represent properties of the individual such as age, gender, home location, the anthrax infection state of the individual, and the ED admission state of the individual. The Bayesian network also contains a global node representing the location of the anthrax release and a global node representing the time of the anthrax release. Temporal information is represented by states of nodes in the network. For example, the global node *Time of Release* has states *never*, *today*, *yesterday*, and *day before yesterday*,

and the person node *Anthrax infection* has states *AAA* (anthrax was absent for the past 3 days), *AAI* (within the past 24 hours the patient was infected with anthrax), *AII* (patient we infected with anthrax between 24 and 48 hours ago and is still infected today), and *III* (patient was infected between 48 and 72 hours ago and is still infected today). Although PANDA models the time period of the outbreak (in days), it does not model a progressive increase in the number of expected outbreak cases over time.

There is a version of PANDA that incorporates the spatial distribution of cases into the model. In this version a node called *Angle of release* is added to the model. This node describes the direction of the airborne anthrax release. The possible values of this node include *north*, *northeast*, *east*, *southeast*, *south*, *southwest*, *west*, and *northwest*. The model defines eight types of rectangular regions centered at the centroid of a zip code. These eight types of regions are also named *north*, *northeast*, *east*, *southeast*, *south*, *southwest*, *west*, and *northwest* respectively according to their direction relative to the centroid of a zip code. For example, assuming the release zip code of an airborne anthrax is 15237 and the value of the *angle of release* node is *northwest*, if the centroid of an individual's home zip code fell in the northwest rectangular region relative to zip code 15237, this individual would be considered to be potentially exposed to anthrax.

2.3.2 Spatial Event Surveillance

2.3.2.1 Methods that Analyze Counts The known methods for spatial event surveillance that analyze counts are the ones that are classified as spatial cluster detection.

Non-Temporal Methods:

Recall that methods for spatial cluster detection attempt to locate spatial subregions of some larger region where the count of occurrences of some event is higher than expected. Non-temporal methods detect clusters based only on the most recent count. Inherent in these methods is that they do not look at patterns of counts over time.

(Kulldorff, 1997, 1999; Kulldorff and Nagarwalla, 1995) developed a classic frequentist non-temporal method called the **spatial scan statistic**, which was implemented in the SaTScan^{TM} software package (Kulldorff, 2004). The scan statistic was first proposed by

(Naus, 1965) as a solution to the multiple hypothesis testing problem. Scan statistics have been used to find clusters of chronic diseases such as breast cancer (Kulldorff et al., 1997) and leukemia (Hjalmars et al., 1996). They have also been used to detect clusters of work-related hazards (Kulldorff et al., 2003) and West Nile virus (Mostashari et al., 2003).

(Jung et al., 2007) developed a version of the spatial scan statistic that considers multinomial variables whose values are ordinal. That is, for each individual in the population it identifies a variable which can have two or more values. It then counts the number of times the variable takes each of its values. We call this a **multinomial** method. This spatial scan is similar to the methods described previously, except in those methods the variable can only have two values. (Kulldorff et al., 2007) developed a version of the spatial scan statistic that considers several counts, each recording occurrences of a different phenomenon. We call this a **multivariate** method.

(Neill et al., 2005a) developed the Bayesian spatial scan statistic. A multivariate version appears in (Neill et al., 2007) and (Neill and Cooper, 2008).

(Neill and Moore, 2004) developed overlapped kd-trees which enable them to speed up the search for clusters over rectangular subregions. (Duczmal and Assuncão, 2004), (Patil and Taillie, 2004), (Kurki and Saarinen, 2006), (Assuncao et al., 2006), (Wieland et al., 2007), and (Duczmal et al., 2008) developed methods for searching arbitrarily shaped subregions, while (Jiang and Cooper, 2007) developed a recursive algorithm that searches over arbitrary subsets of a rectangular grid.

Temporal Methods:

Temporal methods detect clusters based not only on a single set of recent counts, but also on patterns of counts over time.

A spatio-temporal extension of the spatial scan statistic appears in (Kulldorff, 2001) and in (Kulldorff et al., 2005). (Takahashi et al., 2008) developed a flexibly shaped space-time scan statistic. (Neill et al., 2005b) developed a Bayesian expectation-based scan statistic that takes time into account by using historical data to model the expected distribution of counts in each spatial subregion.

2.3.2.2 Entity-Based Methods

PANDA (Cooper et al., 2004), which was discussed in Section 2.3.1.2, can be considered an entity-based, spatio-temporal event surveillance system because it contains nodes for the location and time of the anthrax release. However, it does not attempt to locate a subregion in which an event may be occurring. To my knowledge, the method developed in this thesis is the first method for developing entity-based systems that investigate subregions. However, as will be discussed in Chapter 3, although entity-based systems can be developed using the method in this thesis, it is more general and can be used to develop methods that are not entity-based.

2.3.3 Details of the Spatial Scan Statistics

I will extend PC to both spatial and spatio-temporal systems. In Chapter 5 these systems will be contrasted and compared with the frequentist spatial scan statistic (Kulldorff, 1997, 1999; Kulldorff et al., 2005; Kulldorff et al., 2007), and the Bayesian spatial scan statistic (Neill et al., 2005a; Neill et al., 2007; Neill and Cooper, 2008). So the details of these statistics are presented.

2.3.3.1 The Frequentist Spatial Scan Statistic

When doing spatial cluster detection, we first articulate the subregions of some geographical region G. For example, (Kulldorff, 1997) places a circular window over the region and lets the center of the circle move over the region. For each center, the radius of the circle is varied. (Neill et al., 2005a) represent the entire region by a rectangular grid, and search over rectangular subregions of the grid. Inherent in these methods is that we assume that the entire region G is composed of cells c_i. For example, if we cover G with an $m \times n$ grid, each grid element is a cell. A subregion S of G is the union of any number of cells that form a rectangle. Figure 2.4 illustrates this model.

We are interested in whether some subregion S of G contains a cluster. The null hypothesis H_0 is that there is no cluster, and the alternative hypothesis H_S is that subregion S contains a cluster. (Kulldorff, 1997) developed two different scan statistic models, the Bernoulli model and the Poisson model.

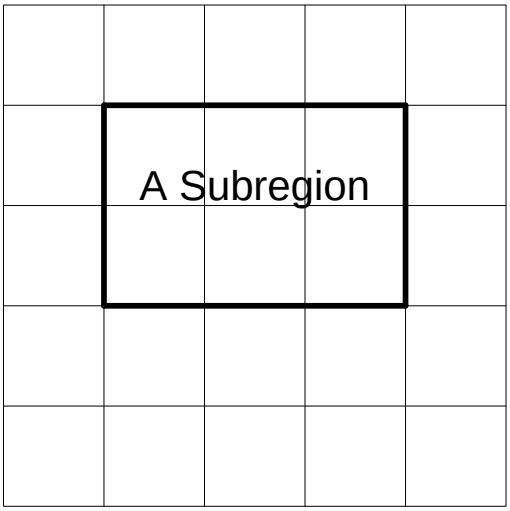


Figure 2.4: An example in which the entire region is covered by a 5×5 grid. One subregion, which is a rectangle, is shown.

In the **Bernoulli spatial scan model**, every cell $c_i \subseteq G$ contains a discrete number of entities. Each entity either does or does not have some property X. For example, the entities could be people, and the property could be that a person visited the Emergency Department with a cough during the past 24 hours. We would be interested if there is a cluster of such entities in some subregion of a particular city or county. As another example, the entities could be stars, and the property could be that the star is a neutron star. We would be interested in whether there is a cluster of neutron stars in some subregion of space. The statements of the hypotheses for the Bernoulli model are as follows:

H_0: For entities in the entire region G, the probability of the entity having property X is q. The event of any one entity having the property is independent of another entity having it.

H_S: For entities in subregion S, the probability of the entity having property X is p. For entities in subregion $G - S$, the probability of the entity having property X is q, where $p > q$. The event of any one entity having the property is independent of another entity having it.

Define the following variables:

B: Total number of entities in G.

$B_{in}^{(S)}$: Total number of entities in S.

$B_{out}^{(S)}$: Total number of entities in $G - S$.

C: Total number of entities in G with property X.

$C_{in}^{(S)}$: Total number of entities in S with property X.

$C_{out}^{(S)}$: Total number of entities in $G - S$ with property X.

(Kulldorff, 1997) shows that the data has highest likelihood under alternate hypothesis H_S for the subregion S that maximizes the following Bernoulli spatial scan statistic:

$$L(S) = \frac{\left(\frac{C_{in}^{(S)}}{B_{in}^{(S)}}\right)^{C_{in}^{(S)}} \left(1 - \frac{C_{in}^{(S)}}{B_{in}^{(S)}}\right)^{B_{in}^{(S)} - C_{in}^{(S)}} \left(\frac{C_{out}^{(S)}}{B_{out}^{(S)}}\right)^{C_{out}^{(S)}} \left(1 - \frac{C_{out}^{(S)}}{B_{out}^{(S)}}\right)^{B_{out}^{(S)} - C_{out}^{(S)}}}{\left(\frac{C}{B}\right)^{C} \left(1 - \frac{C}{B}\right)^{B-C}} \tag{2.1}$$

if

$$\frac{C_{in}^{(S)}}{B_{in}^{(S)}} > \frac{C_{out}^{(S)}}{B_{out}^{(S)}},$$

otherwise

$$L(S) = 1.$$

If subregion S^* has the highest value of the test statistic among all the subregions being tested, we only know that S^* is the most likely subregion to contain a cluster. We determine the statistical significance of the test using a form of Monte Carlo simulation called randomization testing. The technique was originally proposed in (Dwass, 1957), and it was first used in the context of a scan statistic in (Turnball et al., 1990). In this technique, we obtain a large number N of replications of the data set, each of which is generated under the null hypothesis. The p-value of H_S is then equal to

$$\frac{N_{beat} + 1}{N + 1},$$

where N_{beat} is the number of replications in which the subregion with the highest value of the test statistic has a higher value than $L(S^*)$. For example, with 999 such replications, the p-value is .05 if $L(S^*)$ is the 50th highest value obtained relative to the 999 replications.

In the **Poisson spatial scan model**, every cell $c_i \subseteq G$ contains a variable number of points (entities), which we count. For example, the count may be the number of entities with some property. It is assumed that counts are being generated according to an inhomogeneous Poisson process. The statement of the hypotheses are as follows:

H_0: For every cell c_i, the counts are distributed as follows:

$$C_i \sim \text{PoissonDist}(qb_i).$$

where b_i is a baseline count associated with cell c_i.

H_S: For cells $c_i \subset S$, the counts are distributed as follows:

$$C_i \sim \text{PoissonDist}(pb_i),$$

and for cells $c_i \nsubseteq S$, the counts are distributed as follows:

$$C_i \sim \text{PoissonDist}(qb_i),$$

where $p > q$.

Define the following variables:

$$B = \sum_i b_i.$$
$$B_{in}^{(S)} = \sum_{i:c_i \subseteq S} b_i.$$
$$B_{out}^{(S)} = \sum_{i:c_i \nsubseteq S} b_i.$$
$$C = \sum_i c_i.$$
$$C_{in}^{(S)} = \sum_{i:c_i \subseteq S} c_i.$$
$$C_{out}^{(S)} = \sum_{i:c_i \nsubseteq S} c_i.$$

(Kulldorff, 1997) shows that the most significant subregion S is the one that maximizes the following Poisson spatial scan statistic:

$$F(S) = \frac{\left(\frac{C_{in}^{(S)}}{B_{in}^{(S)}}\right)^{C_{in}^{(S)}} \left(\frac{C_{out}^{(S)}}{B_{out}^{(S)}}\right)^{C_{out}^{(S)}}}{\left(\frac{C}{B}\right)^{C}} \tag{2.2}$$

if

$$\frac{C_{in}^{(S)}}{B_{in}^{(S)}} > \frac{C_{out}^{(S)}}{B_{out}^{(S)}},$$

otherwise

$$F(S) = 1.$$

As is the case for the Bernoulli spatial scan statistic, we determine the statistical significance of our finding by doing randomization testing using a large number of replications of the data set.

2.3.3.2 The Temporal and Multivariate Frequentist Scan Statistics A temporal version of the spatial scan statistic appears in (Kulldorff, 2001) and (Kulldorff et al., 2005). It is much like the basic version described above except that instead of a circular window in two dimensions, the space-time scan statistic searches over cylindrical windows in three dimensions. The base of the cylinder represents space in the same way as the basic scan statistic, while the height of the cylinder represents time. When a disease outbreak occurs, the counts ordinarily increase (with daily fluctuation) until some peak is reached, and then decline. Kulldorff's temporal spatial scan statistic does not consider this phenomenon. (Neill et al., 2005b) developed a temporal version of the spatial scan statistic that does model increasing counts. This version looks at counts from previous days, and, in the alternative hypothesis, uses a different parameter p for each day, where the values of these parameters are assumed to be monotonically increasing. For example, suppose we are looking at data from day_1, day_2, day_3 and day_4, where day_4 is today, day_3 is yesterday, day_2 is two days ago, and day_1 is three days ago. It is assumed that the count for a cell $c_i \subset S$ on day_1 are distributed PoissonDist($p_1 b_i$), the count on day_2 is distributed PoissonDist($p_2 b_i$), and so on. It is further assumed that the sequence $[p_1, p_2, p_3, p_4]$ is monotonically increasing.

(Kulldorff et al., 2007) developed a multivariate version of the spatial scan statistic that considers several counts, each recording occurrences of a different type of event. For example, we may be looking at over-the-counter (OTC) sales of different products. One type of event (product) could be the purchase of antidiarrheal medication, a second type could be the purchase of cold/cough medication, and a third type could be the purchase of a thermometer. A count would be recorded for each of these three types of events. The total log likelihood in this multivariate version is the sum of the individual log likelihoods for each separate event type (Equation 2.1 or 2.2).

All these spatial scan statistics have been implemented in the SaTScan^{TM} software package, which is available for free at http://www.satscan.org/. Methods available there include the basic spatial scan statistic, a temporal spatial scan statistic, a multivariate spatial scan statistic, and a spatial scan statistic which is both temporal and multivariate.

2.3.3.3 The Bayesian Spatial Scan Statistic There are several problems with the frequentist spatial scan statistic. First, it is quite difficult to incorporate any prior information about the outbreak. For example, it would be hard to use our prior beliefs about the size of a possible outbreak and its impact on our observed counts. Second, its accuracy depends on the correctness of our maximum likelihood parameter estimates. This means the model is prone to parameter overfitting, and therefore may lose detection power because of model misspecification. Finally, the statistic is computationally very costly due to the necessity to create replica grids.

(Neill et al., 2005a) remedied these problems by developing a Bayesian spatial scan statistic. Their Bayesian model uses prior information about the likelihood, size, and impact of an outbreak. If these priors are well-chosen, it should exhibit better detection performance that the frequentist model. Second, the Bayesian approach averages over possible values of the model parameters rather than using maximum likelihood estimates of these parameters. This approach is less prone to overfitting. Finally, in this approach there is no need to create replica grids.

As in the frequentist approach, the Bayesian approach compares the null hypothesis H_0 with the alternative hypotheses H_S, each of which represents that there is a cluster in

subregion S. This approach makes the following assumptions, which are like those in the frequentist approach (the variables used here are the same as those defined in Section 2.3.3.1):

H_0: For every cell c_i, the counts are distributed as follows:

$$C_i \sim \text{PoissonDist}(q_{all} b_i).$$

where b_i is a baseline count associated with cell c_i.

H_S: For cells $c_i \subset S$, the counts are distributed as follows:

$$C_i \sim \text{PoissonDist}(q_{in} b_i),$$

and for cells $c_i \nsubseteq S$, the counts are distributed as follows:

$$C_i \sim \text{PoissonDist}(q_{out} b_i).$$

The difference between the Bayesian approach and the frequentist approach is that this approach assumes a hierarchical Bayesian model in which the disease rates q and p are themselves drawn from Gamma distributions. Specifically, this approach assumes the following:

H_0: For every cell c_i, the disease rate are distributed as follows:

$$q_{all} \sim \text{GammaDist}(\alpha_{all}, \beta_{all}).$$

H_S: For cells $c_i \subset S$, the disease rates are distributed as follows:

$$q_{in} \sim \text{GammaDist}(\alpha_{in}, \beta_{in})$$

and for cells $c_i \nsubseteq S$, the disease rates are distributed as follows:

$$q_{out} \sim \text{GammaDist}(\alpha_{out}, \beta_{out}).$$

(Neill et al., 2005a) discuss how to obtain the parameters α and β.

Based on this model, (Neill et al., 2005a) derive the following likelihoods (the variables used here are the same as those defined in Section 2.3.3.1):

$$P(Data|H_0) = \frac{(\beta_{all})^{\alpha_{all}} \Gamma(\alpha_{all} + C)}{(\beta_{all} + B)^{\alpha_{all}+C} \Gamma(\alpha_{all})}$$

$$P(Data|H_S) = \frac{\left(\beta_{in}^{(S)}\right)^{\alpha_{in}^{(S)}} \Gamma\left(\alpha_{in}^{(S)} + C_{in}^{(S)}\right)}{\left(\beta_{in}^{(S)} + B_{in}^{(S)}\right)^{\alpha_{in}^{(S)}+C_{in}^{(S)}} \Gamma\left(\alpha_{in}^{(S)}\right)} \times \frac{\left(\beta_{out}^{(S)}\right)^{\alpha_{out}^{(S)}} \Gamma\left(\alpha_{out}^{(S)} + C_{out}^{(S)}\right)}{\left(\beta_{out}^{(S)} + B_{out}^{(S)}\right)^{\alpha_{out}^{(S)}+C_{out}^{(S)}} \Gamma\left(\alpha_{out}^{(S)}\right)}$$

The Bayesian spatial scan statistic is computed by calculating $P(Data|H_S)P(H_S)$ for each subregion S, and then using Bayes' Theorem to compute

$$P(H_S|Data) = \frac{P(Data|H_S)P(H_S)}{\sum_R P(Data|H_R)P(H_R)}.$$

We see that since this approach determines posterior probabilities, there is no need to use randomization to create replications of the data set and to determine significance, which greatly decreases its time complexity relative to the frequentist spatial scan statistic.

2.3.3.4 The Multivariate Bayesian Scan Statistic A multivariate version of the Bayesian spatial scan statistic appears in (Neill et al., 2007) and (Neill and Cooper, 2008). They combine the prior probability of an outbreak in each spatial region with the likelihood of the multivariate data using Bayes' Theorem. To compute the data likelihood given the either the null or alternative hypothesis, they use a Gamma-Poisson model (as in the univariate Bayesian spatial scan statistic) for each event type. They assumed that the counts for each event type are conditionally independent given the outbreak type, affected region, and outbreak parameters. The parameter priors for each event type are learned from the time series of past counts.

3.0 METHODOLOGY

This chapter describes Bayesian network architectures for event surveillance, spatial event surveillance, temporal event surveillance, and spatio-temporal event surveillance. Each architecture is an enhancement of the previous one. The architectures are created in sequence:

1. An event surveillance architecture.
2. A spatial event surveillance architecture.
3. A temporal event surveillance architecture.
4. A spatio-temporal event surveillance architecture.

3.1 THE BAYESNET CLASS OF EVENT SURVEILLANCE MODELS

This section presents a description of the high-level Bayesian network architecture representing the BayesNet class of event surveillance models and then gives several concrete examples.

3.1.1 The High-Level Bayesian Network Architecture

Suppose we are investigating whether there is an event of interest in some region. Let E be a random variable whose value is "yes" if the event of interest occurred or is occurring, and whose value is "no" otherwise. Besides the variable E, there can be a set of attribute variables which represent properties of the event of interest, a set of intermediate variables which depend on the properties of the event of interest, and a set of observable variables which depend on the intermediate variables. These observable variables comprise our $Data$. Figure 3.1 shows a high-level Bayesian network architecture representing this class of models. Any

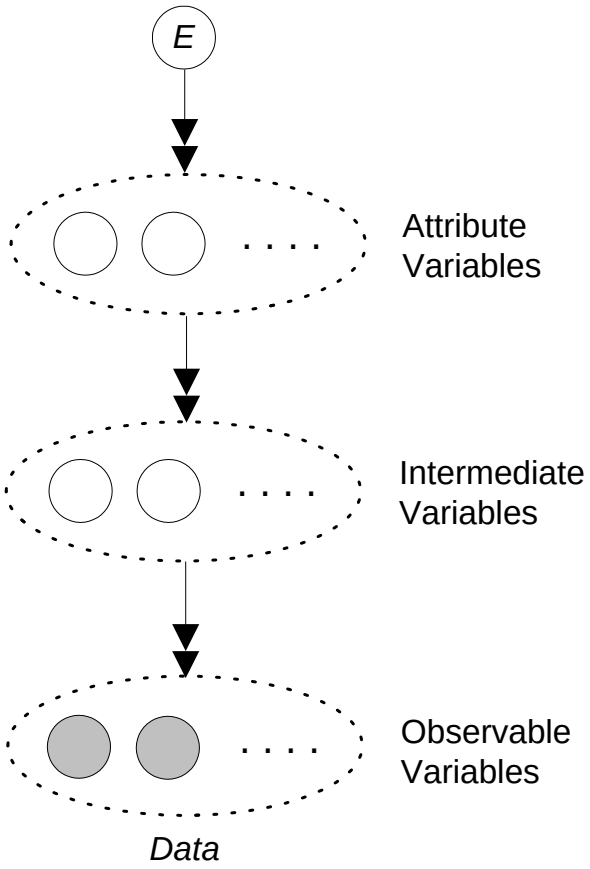


Figure 3.1: The high-level BayesNet Bayesian network architecture. The value of E is "yes" if the event of interest occurred, and is "no" otherwise. The sets of variables enclosed by ovals represent Bayesian subnetworks. The attribute variables are properties of the event of interest, the intermediate variables depend on the properties of the event of interest, and the observable variables depend on the intermediate variables. The shaded observable variables are the measured variables and comprise our $Data$. The unshaded variables are unmeasured. The double arrowed edges indicate that there can be more than one edge from each variable in a given set to the variables in the set below it. In general, there need not be any attribute or intermediate variables.

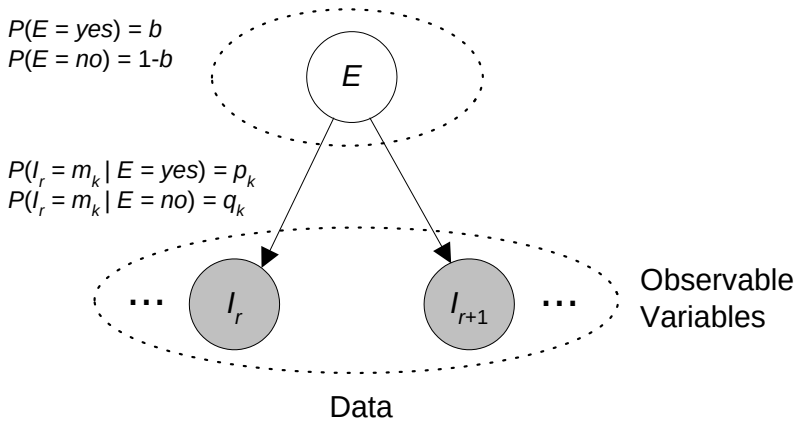


Figure 3.2: A simple example of a BayesNet model.

model in this class is called a **Bayesian Network (BayesNet) model.** If each intermediate variable represents an individual in a population, and there is a set of observable variables for each such individual, it would be an entity-based model. In this thesis only BayesNet models that are entity-based will be considered. However, the theory does not require that they be entity-based. For example, suppose E represents the occurrence of an influenza outbreak, and the only observable variables is C, which is the count of OTC sales of thermometers. The variable C depends on E, and we can model this dependency using the DAG $E \rightarrow C$. This is a BayesNet model containing no attribute or intermediate variables and which is not entity-based.

In a non-spatial, non-temporal model, the data are obtained from the entire region being monitored, and new data are obtained each day (or at whatever our time unit may be). The Bayesian network is used to compute

$$P(E = yes|Data).$$

3.1.2 A Simple Example of a BayesNet Model

3.1.2.1 The Model Figure 3.2 shows a simple example of a BayesNet model, which has no global or intermediate variables. For the sake of concreteness, let us give the variables

meaning. Suppose that the variable E has value "yes" if there is currently an outbreak of influenza and the value "no" otherwise. There is a variable I_r for each individual r in the entire region G. So this is an entity-based system. There are no variables describing properties of the event per se (beyond data about entities in the population) and no intermediate variables. The possible values of I_r are our manifestations m_k for each individual. In this example, suppose that they are the chief complaints with which the individual might present in the Emergency Department, where one value is "noED", which means the individual did not visit the Emergency Department. Other possible chief complaints include cough, and fever/chills. Note that $I_r = m_k$ is an assignment of chief complaint m_k for individual I_r. The following is a concrete example.

Example 3.1. *The Bayesian network in Figure 3.2 could have these chief complaints and probability distributions:*

$$\begin{aligned} m_1 &= cough \\ m_2 &= fever/chills \\ m_3 &= noED \end{aligned}$$

$$P(E = yes) = 0.0001$$

$$\begin{aligned} P(I_r &= cough|E = yes) = 0.001 \\ P(I_r &= fever/chills|E = yes) = 0.002 \\ P(I_r &= noED|E = yes) = 0.997 \end{aligned}$$

$$\begin{aligned} P(I_r &= cough|E = no) = 0.0003 \\ P(I_r &= fever/chills|E = no) = 0.0005 \\ P(I_r &= noED|E = no) = 0.9992. \end{aligned}$$

3.1.2.2 The Inference Algorithm The *Data* consists of the values of I_r for all individuals r in region G. Since there could be thousands, or even millions, of individuals in G, we would not explicitly construct the Bayesian network in Figure 3.2, and instantiate I_r for all r. Rather, due to the fact that the Bayesian network structure entails that individuals' chief complaints are conditionally independent given the value of E, we can compute the likelihoods of the data as follows:

$$P(Data|E = yes) = \prod_k (p_k)^{C_k}$$

$$P(Data|E = no) = \prod_k (q_k)^{C_k},$$

where C_k is the number of individuals with the kth chief complaint, and p_k and q_k are defined in Figure 3.2. Then using Bayes' Theorem, we compute that

$$P(E = yes|Data) = \frac{P(Data|E = yes)P(E = yes)}{P(Data|E = yes)P(E = yes) + P(Data|E = no)P(E = no)}.$$

3.1.3 PANDA-CDCA

I now describe a more complex example of a BayesNet model, namely the Bayesian network in PANDA-CDCA (PC). This system was developed in (Cooper et al., 2007) and not in this thesis. However, it is discussed in detail because it was used as the basis for the domain models that are developed and evaluated in this dissertation.

3.1.3.1 The Model Figure 3.3 shows the Bayesian network in PC. There are nodes D_r and I_r corresponding to each individual r in the population. In Figure 3.3 I denoted these multiple nodes by showing a couple of them. An alternative representation is to use a **plate** as described in (Buntine, 1994) and (Spiegelhalter, 1998). This representation appears in Figure 3.4. The box (plate) around the subgraph $D \rightarrow I$ indicates that this subgraph is repeated N times, where N is the number of individuals in the population.

Each node in the network is described next.

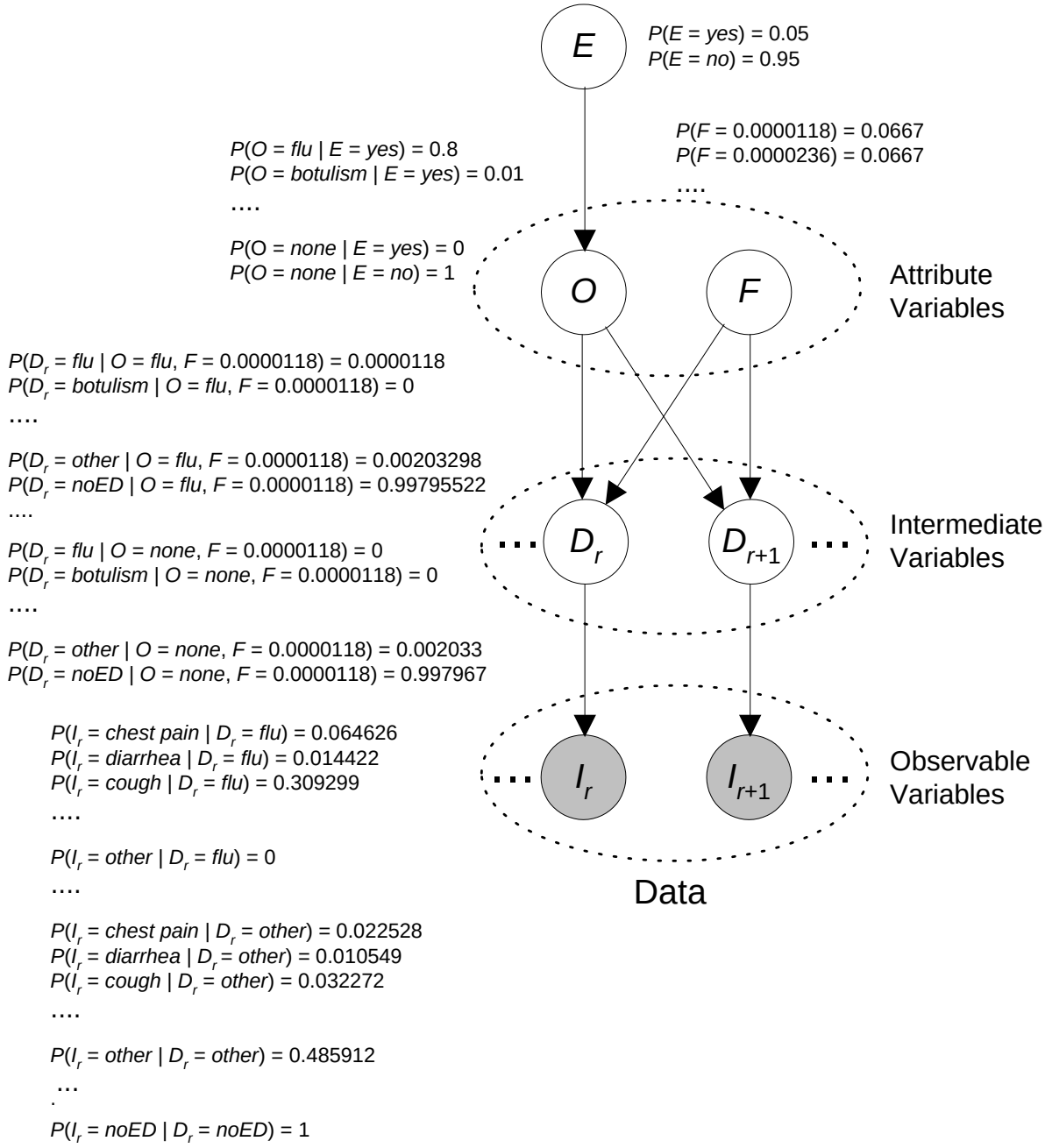


Figure 3.3: The PC Bayesian network. See the text for a description of the variables.

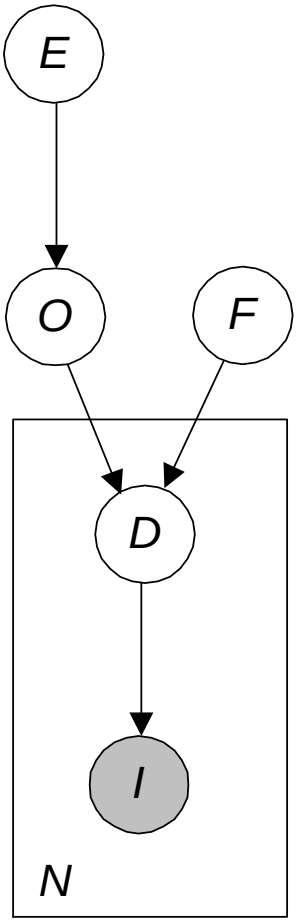


Figure 3.4: The Bayesian network in PC using a plate to represent multiple occurrences of the subgraph $D \rightarrow I$. This subgraph is repeated N times where N is the number of individuals in the population.

E: This node represents whether there is an ongoing outbreak. The value "yes" represents that there is an ongoing outbreak of one of the outbreak diseases represented by O during the previous 24-hour period.

Note that $P(E = yes) = 0.05$. This prior probability is fairly high because influenza is one of the diseases being modeled, and it is relatively common during any given year. PC does not currently model different seasons of the year, and thus, this probability is a constant throughout the year.

O: This node represents which outbreak disease is occurring if there is an outbreak. The prior probabilities for variable O were assessed by the project's infectious disease expert, Dr. John Dowling, based on the literature and subjective estimates. There are 13 possible outbreak diseases, two of which are shown in Figure 3.3 (influenza and botulism). The possible outbreak diseases include the following CDC Category A diseases:

1. anthrax stage 1
2. anthrax stage 2
3. plague stage 1
4. plague stage 2
5. smallpox
6. tularemia
7. botulism
8. marburg hemorrhagic fever stage 1
9. marburg hemorrhagic fever stage 2.

The CDC Category A diseases can be either easily transmitted from person to person or easily disseminated, result in high mortality rates with a potential for major public health impact, might cause public panic and social disruption, and require special action for public health preparedness.

The other possible outbreak diseases included are the following ones that may be confused with CDC Category A diseases:

1. flu

2. *Cryptosporidium*

3. hepatitis A.

The 13th value of O is "none", which represents a population-disease state in which there is no outbreak disease. Note that $P(O = none \mid E = \text{no}) = 1.0$. Thus, each of the outbreak diseases listed above has a probability of 0, when $E = \text{no}$.

PC assumes that outbreak diseases are mutually exclusive. For example, it assumes there would not be influenza and botulism outbreaks occurring simultaneously. Although in reality different outbreaks could occur concurrently, this event is unlikely, and therefore the model currently assumes it does not happen.

F: The value of this node is itself a probability. This node represents the probability of an individual both being afflicted with the outbreak disease and going to the ED, given that an outbreak is ongoing. For computational efficiency reasons, the states of this node were discretized into 15 numerical values, two of which are shown in Figure 3.3. This node indicates the extent of the outbreak, if one is occurring. Since the value f of F is a probability, the probability distribution of F is a higher order probability distribution.

The population being monitored by the EDs for which PC obtains data is estimated to be 423,076 people. This number is 29% of the 1,458,883 people who live in Allegheny County, according to the 2000 US census. The value of 29% is an estimate of the ED coverage of Allegheny County represented in the biosurveillance database that I used.

The possible values of F correspond to expected number of outbreaks cases of 5, 10, 15, 20, 25, 50, 75, 100, 125, 150, 175, 200, 225, 250, and 275 according to the following calculations. The mean number of ED cases per day when there is not an outbreak is estimated to be 577 according to the 2004 biosurveillance data, and the standard deviation is about $\sigma = 54$. It was assumed that the expected value of the increased number of ED cases during an outbreak ranged between $0.1\sigma = 0.1 \times 54 \approx 5$ to $5\sigma = 5 \times 54 \approx 275$. The 15 values above were then taken from this range. Finally, the values of F were obtained by dividing these numbers by 423,076, which is the number of people

in the population. For example,

$$\frac{5}{423,076} = 1.18 \times 10^{-5}.$$

In general, there would be an edge from O to F, which allows the distribution of F to depend on which outbreak disease (if any) is occurring in the population. That is, some outbreak diseases might be expected to affect a larger fraction of the population than other outbreak diseases. However, the current 24 hours of data considered by PC can represent any stage of a disease outbreak. Earlier stages would tend to have lower values of F, whereas later stages would tend to have higher values of F. Given the uncertainty of the disease stage, PC currently does not include an arc from O to F, because for any given disease many different distributions of F are possible, depending on the stage.

D_r: This node represents the ED disease state of the rth individual. There is one such node for each individual r in the population. So this is an entity-based system. The value "noED" means the individual does not visit the ED. The value "other" means the individual arrives in the ED only with some non-outbreak disease (e.g., a broken arm). The value "anthrax", for example, means the individual arrives in the ED with an anthrax infection. This means that the individual arrived with anthrax due to an outbreak, and not due to a natural, sporadic anthrax. An individual presenting with sporadic anthrax would be classified as having value "other" disease. The same holds for the other outbreak diseases.

The probabilities for node D_r were obtained as follows. If there is no outbreak occurring in the population, it is assumed the individual could not have an outbreak disease. Therefore, when there is no outbreak, the individual could arrive in the ED only with a non-outbreak disease. The probability of this event is called p_{other}. So when there is no outbreak, the probability of not visiting the ED is $1 - p_{other}$. These probabilities are estimated using the ED data from the previous year. In this dissertation, I use data from 2005 for testing outbreak detection performance, as described in Chapter 4. Therefore,

I use data from 2004 for estimating model parameters. In 2004 in Allegheny County probability estimates for PC are as follows:

$$\begin{aligned} P(D_r = other|O = none, F = f) &= p_{other} \\ &= 0.002033 \end{aligned}$$

$$\begin{aligned} P(D_r = noED|O = none, F = f) &= 1 - p_{other} \\ &= 0997967. \end{aligned}$$

The probabilities of arriving in the ED with outbreak diseases given there is an outbreak of disease d is based on the value of F as follows:

$$P(D_r = d|O = d, F = f) = f$$

$$P(D_r = c|O = d, F = f) = 0 \qquad \text{for } c \neq d.$$

It is assumed that the factors that inhibit an individual from going to the ED with an outbreak disease act independently of the factors that inhibit the individual from going to the ED with a non-outbreak disease. Accordingly, a noisy-OR relationship (Pearl, 1988) is used as follows:

$$P(D_r = noED|O = d, F = f) = (1 - p_{other})(1 - f).$$

Finally,

$$\begin{aligned} &P(D_r = other|O = d, F = f) \\ &= 1 - P(D_r = d|O = d, F = f) - P(D_r = noED|O = d, F = f) \\ &= 1 - f - (1 - p_{other})(1 - f) \\ &= p_{other}(1 - f). \end{aligned}$$

I_r: This node represents each of the possible chief complaints individual r could have when arriving in the ED. There are 54 possible chief complaints, one of which is "other", which means the chief complaint was not one of the 53 specific chief complaints represented in the network. The 55th value of the node is "noED", which means the individual did not visit the ED and thus did not have a literal chief complaint.

Typically, a patient can have many findings. The PC model, however, needs the probability that a finding is a chief complaint, because for patients who come to the ED, the chief complaint is typically the only clinical information that is given to PC. The probability of a patient having a finding (e.g., a finding of fever) and having the finding be his or her chief complaint (e.g., a chief complaint of fever) are in general not the same. A chief-complaint probability is almost always lower. In PC, the finding probabilities were provided by an infectious disease expert, and a method was used by the PC developers to transform finding probabilities to chief complaint probabilities.

3.1.3.2 Mapping Chief Complaint Data ED chief complaints are a popular data source used by many biosurveillance related systems due to their timeliness and availability (Chapman et al., 2004; Chapman et al., 2005; Espino et al., 2001). PC uses ED chief complaints data as its input. A chief complaint is ordinarily a free-text phrase entered by triage personnel. PC maps such a phrase to one of the 54 chief complaints it represents. First, I briefly review methods for handling chief complaint data. Then I describe how PC maps a phrase to a chief complaint.

The methods for handling chief complaint data can be categorized into two main categories: the rule-based approach and the probabilistic approach. The rule-based approach often consists of two steps. In the first step, the chief complaint records are converted to a group of predefined terms by performing either a table lookup or keyword match. In the second step, the groups of predefined terms are translated into final syndromic categories using a set of rules. With the rule-based approach it is difficult to handle new symptoms that are not previously encountered. Examples of the rule-based approach include EARS (Hutwagner et al., 2003), ESSENCE (Lombardo et al., 2003), DOHMH syndromic coding system (Mikosz et al., 2004), and an ontology-enhanced approach discussed in (Lu et al.,

2008). Both the EARS and the DOHMH systems use the keyword matching scheme and a set of predefined mapping rules. The ESSENCE system uses a weighted key word matching scheme. It treats each chief complaint as a document and it runs a query against documents. In (Lu et al., 2008), an ontology-enhanced method that basically follows a rule-based design was introduced. This method focuses on automatically expanding the coverage of the symptom lookup table by exploiting the semantic relations between symptoms.

The probabilistic approach often uses a Bayesian network model to classify chief complaint records. In this approach, a Bayesian network model is used to identify the most likely syndromic category for a given a chief complaint record through its built-in inference engine. Examples of the probabilistic approach include CoCo naïve Bayesian classifier (Tsui et al., 2003; Olszewski, 2003) and the Medical Probabilistic Language Understanding System (M+) introduced in (Chapman et al., 2005). M+ is a robust chart-based syntactic parser with a Bayesian network-based semantic model for extracting information from text records. M+ has been applied to the fields of chest radiography and brain CT scans. In (Chapman et al., 2005) M+ was trained to perform the task of free-text triage chief complaints classification.

The rule-based approach is used in PC. I describe the details of it next. In what follows I call the chief complaints in PC "findings" so as to distinguish them from the actual free-text chief complaint strings. A table was constructed that lists each finding with a set of search phrases. Both the findings and the search phrases were developed with the guidance of an infectious disease expert. The following is a portion of the table:

Finding	Search Phrases
difficult swallowing	difficult swallowing, swallowing difficulty, dysphagia
dyspnea	dyspnea, shortness of breath, sob
fever/chills	fever, chills, high temperature
insomnia	insomnia, cannot sleep, difficulty sleep
nausea/vomiting	nausea, vomit

For each individual's chief complaint string, the algorithm visits the findings in sequence, and checks whether a search phrase, which is associated with a finding, is contained in the chief complaint string. When a search phrase is found to be contained in a chief complaint string, we say that a match is found. When a match is found, the algorithm assigns the

finding associated with the search phrase as the individual's chief complaint. For example, if the chief complaint is "difficulty sleeping," the search phrase "difficulty sleep" should be a matching phrase when the "insomnia" finding was visited, and so "insomnia" would be assigned as the individual's chief complaint. If no match is found for any finding, the individual's chief complaint is assigned the value "other_finding". Note that if more than one finding has a search phrase contained in the chief complaint string, only the first finding encountered will be assigned.

3.1.3.3 The Inference Algorithm A large Bayesian network containing a node for each patient in the population exists conceptually, but we do not need to actually create and perform inference using an explicit Bayesian network. Rather, to do inference with the network, we can proceed as follows. On each day, we know the value of I_r for each individual r in the population. The set of all these values is our $Data$.

First, we have that[1]

$$\begin{aligned} P(Data|E = no) &= P(Data|O = none) \\ &= \prod_k (P(I_r = m_k|O = none))^{C_k}, \end{aligned} \tag{3.1}$$

where C_k is the number of individuals with the kth chief complaint m_k. Note that one of the chief complaints is "noED", which means the individual did not visit the ED. The reason we can compute $P(Data|O = none)$ by multiplying the individual probabilities is that the nodes labeled I_r are conditionally independent given values of O and F, and by construction F's value is irrelevant when $O = none$.

The value of $P(I_r = m_k|O = none)$ for each patient who went to the ED could be computed by performing inference using the Bayesian network in Figure 3.3. However, we can obtain this value more efficiently as follows:

[1]Note that $E = no$ and $O = none$ designate the same event of there being no outbreak in the population.

$$
\begin{aligned}
P(I_r = m_k|O = none) &= \sum_c P(I_r = m_k|D_r = c)P(D_r = c|O = none) \\
&= P(I_r = m_k|D_r = other)P(D_r = other|O = none) + \\
&\quad P(I_r = m_k|D_r = noED)P(D_r = noED|O = none) \\
&= P(I_r = m_k|D_r = other) \times p_{other} + \\
&\quad P(I_r = m_k|D_r = noED) \times (1 - p_{other}). \qquad (3.2)
\end{aligned}
$$

Next, we have that

$$P(Data|O = d) = \sum_f P(Data|O = d, F = f)P(F = f). \qquad (3.3)$$

To obtain the terms in the expression on the right of Equation 3.3, we have that

$$P(Data|O = d, F = f) = \prod_k (P(I_r = m_k|O = d, F = f))^{C_k}. \qquad (3.4)$$

where C_k is the number of individuals with the kth chief complaint m_k.

Again, rather than performing direct inference using a Bayesian network, we proceed as follows.

$$
\begin{aligned}
&P(I_r = m_k|O = d, F = f) \\
&= \sum_c P(I_r = m_k|D_r = c)P(D_r = c|O = d, F = f) \\
&= P(I_r = m_k|D_r = d)P(D_r = d|O = d, F = f) + \\
&\quad P(I_r = m_k|D_r = other)P(D_r = other|O = d, F = f) + \\
&\quad P(I_r = m_k|D_r = noED)P(D_r = noED|O = d, F = f) \\
&= P(I_r = m_k|D_r = d)f + P(I_r = m_k|D_r = other)p_{other}(1 - f) + \qquad (3.5) \\
&\quad P(I_r = m_k|D_r = noED)(1 - p_{other})(1 - f).
\end{aligned}
$$

Using Bayes' Theorem,

$$P(O = d|Data) = \frac{P(Data|O = d)P(O = d)}{\sum_c P(Data|O = c)P(O = c)}. \qquad (3.6)$$

The prior probability of an outbreak disease is computed as follows:

$$\begin{aligned} P(O = d) &= P(O = d|E = yes)P(E = yes) + P(O = d|E = no)P(E = no) \\ &= P(O = d|E = yes)P(E = yes). \end{aligned}$$

Finally, the probability of an outbreak is given by

$$P(E = yes|Data) = \sum_{d \neq none} P(O = d|Data). \tag{3.7}$$

3.1.3.4 A Time Complexity Analysis of the Inference Algorithm Let us first analyze the time complexity of computing the value of $P(E = yes|Data)$ in Equation 3.7. To compute this value, we must do the calculations in Equations 3.1 through 3.6.

First define the following variables:

N_C: Number of chief complaints.

N_F: Number of values of F.

N_D: Number of outbreak diseases.

To do the calculation in Equation 3.1, N_C computations of Equation 3.2 are required. The calculation in Equation 3.2 requires constant time. The calculation in Equation 3.3 requires N_F computations of Equation 3.4. Each computation of Equation 3.4 requires N_C computations of Equation 3.5. The calculation in Equation 3.5 requires constant time. Equation 3.6 must be computed for every outbreak disease d, which means the computation of Equation 3.3 must be done N_D times. Finally, we must sum over all outbreak diseases to obtain the denominator in Equation 3.6, compute Equation 3.6 for all outbreak diseases, and sum over all outbreak diseases in Equation 3.7. Therefore, in total the time to compute $P(E = yes|Data)$ in Equation 3.1 is

$$N_C + N_C \times N_F \times N_D + 3N_D \in \theta(N_C \times N_F \times N_D).$$

In deriving this result, we have assumed that the probabilities in the Bayesian network can be obtain in $\theta(1)$ time by a table look-up.

The algorithm has constant running time relative to the number of individuals in the population. However, we need to pre-process the data to obtain the values of C_k for every chief complaint m_k. This calls for checking a record for every individual who visited the ED. Let N_{ED} be the number of individuals who visited the ED. So the overall running time of the algorithm is

$$\theta(N_{ED} + N_C \times N_F \times N_D).$$

3.2 THE BAYESNET-S CLASS OF SPATIAL EVENT SURVEILLANCE MODELS

This section presents a description of the high-level Bayesian network architecture representing the BayesNet-S class of spatial event surveillance models and then gives several concrete examples.

3.2.1 The High-Level Bayesian Network Architecture

Let G be the spatial region we are monitoring. We start with the high-level Bayesian network architecture in Figure 3.1. Then one additional random variable SUB is added to the set of attribute variables. The value of SUB is S_j if there is an event of interest in subregion S_j and is "none" if there is no event of interest in any subregion. Recall from Section 2.3.3.1 that a subregion is the union of any number of cells. In the applications considered here, G is covered with an $n \times n$ grid, each grid element is a cell, and only subregions that are rectangles are considered.

A high-level Bayesian network architecture representing this class of models appears in Figure 3.5. Any model in this class is called a **Bayesian Network Spatial (BayesNet-S) model.**

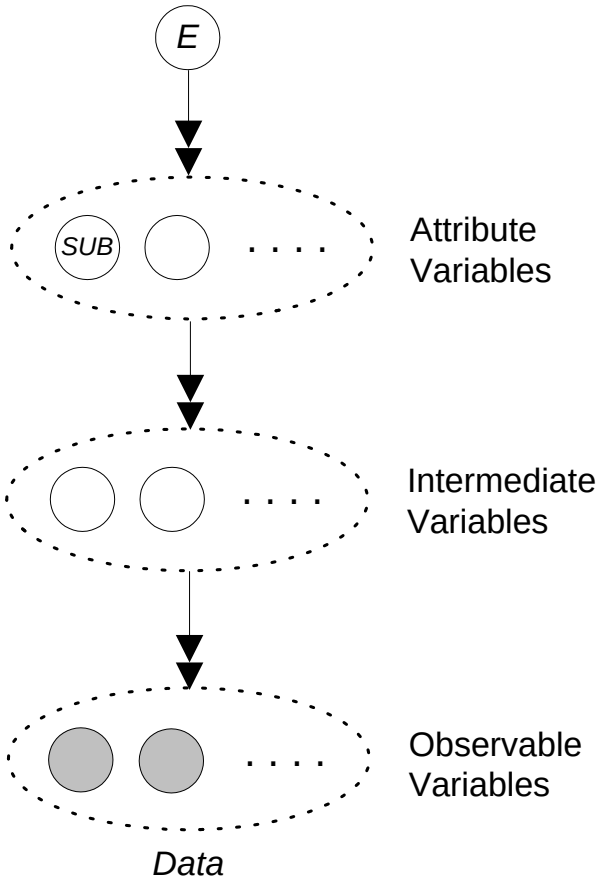


Figure 3.5: The high-level BayesNet-S Bayesian network architecture. The discussion in the caption of Figure 3.1 pertains to this figure. There is always one attribute variable SUB, whose value is the subregion in which the event is occurring if there is an event.

3.2.2 A Simple Example

This section provides a simple example of a BayesNet-S model, and this instance is used to show that, in a sense, the frequentist spatial scan statistics (see Section 2.3.3.1) are special cases of BayesNet-S.

3.2.2.1 The Model Consider the BayesNet-S model in Figure 3.6. It is based on a Bayesian network like the one in Figure 3.2, except each observable variable has only two possible values. The value X means an individual has property X and the value noX means the individual does not. If there is an event (E = yes), the probability of a subregion S_j containing the event is b (i.e. a uniform distribution over the subregions is assumed), while if there is no event (E = no), the probability that there is no event in any subregion is 1. For each individual r, there is a location variable Loc_r, whose value is known at run-time and which represents the individuals home location such as zip code. For each value of Loc_r and each value S_j of SUB, we need to know whether or not $Loc_r \in S_j$.

3.2.2.2 The Inference Algorithm Define the following variables, which are the same variables defined in Section 2.3:

B: Total number of individuals in region G.

$B_{in}^{(S_j)}$: Total number of individuals in S_j.

$B_{out}^{(S_j)}$: Total number of individuals in $G - S_j$.

C: Total number of individuals in G with property X.

$C_{in}^{(S_j)}$: Total number of individuals in S_j with property X.

$C_{out}^{(S_j)}$: Total number of individuals in $G - S_j$ with property X.

$Data_{in}^{(S_j)}$: The data on individuals in S_j.

$Data_{out}^{(S_j)}$: The data on individuals in $G - S_j$.

Then from the Bayesian Network in Figure 3.6 we see that

$$P(Data|SUB = none) = q^C(1-q)^{B-C},$$

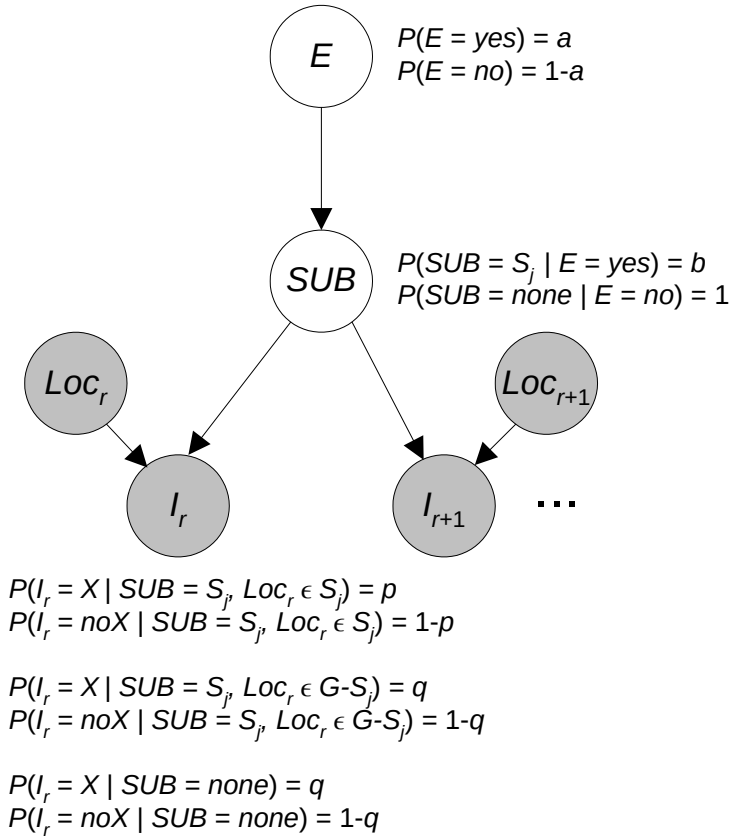


Figure 3.6: A simple example of a BNetScan-S model.

and for subregion S_j that is hypothesized to contain an outbreak we have the following:

$$P(Data_{in}^{(S_j)}|SUB = S_j) = p^{C_{in}^{(S_j)}}(1-p)^{B_{in}^{(S_j)}-C_{in}^{(S_j)}}$$

$$P(Data_{out}^{(S_j)}|SUB = S_j) = q^{C_{out}^{(S_j)}}(1-q)^{B_{out}^{(S_j)}-C_{out}^{(S_j)}}$$

$$\begin{aligned} P(Data|SUB = S_j) &= P(Data_{in}^{(S_j)}|SUB = S_j)P(Data_{out}^{(S_j)}|SUB = S_j) \\ &= p^{C_{in}^{(S_j)}}(1-p)^{B_{in}^{(S_j)}-C_{in}^{(S_j)}} q^{C_{out}^{(S_j)}}(1-q)^{B_{out}^{(S_j)}-C_{out}^{(S_j)}}. \end{aligned} \tag{3.8}$$

We then use Bayes' Theorem to compute

$P(SUB = S_j|Data)$

$$= \frac{P(Data|SUB = S_j)P(SUB = S_j)}{\sum_i P(Data|SUB = S_i)P(SUB = S_i)}$$

$$= \frac{p^{C_{in}^{(S_j)}}(1-p)^{B_{in}^{(S_j)}-C_{in}^{(S_j)}} q^{C_{out}^{(S_j)}}(1-q)^{B_{out}^{(S_j)}-C_{out}^{(S_j)}} P(S_j)}{\sum_{S_i \neq none} p^{C_{in}^{(S_i)}}(1-p)^{B_{in}^{(S_i)}-C_{in}^{(S_i)}} q^{C_{out}^{(S_i)}}(1-q)^{B_{out}^{(S_i)}-C_{out}^{(S_i)}} P(S_i) + q^C(1-q)^{B-C}P(none)}.$$

The probability that there is an event in some subregion is equal to

$$\sum_{S_i \neq none} P(SUB = S_i|Data).$$

3.2.2.3 The Spatial Scan Statistic Method as a Special Case In a sense, the frequentist spatial scan statistics are special cases of the model just developed. I will now show this. Suppose that our estimates of the conditional probabilities in Equation 3.8 are based on the fraction of people who showed the respective behaviors in the subregions. We then have

$$P(Data|SUB = S_j)$$

$$= \left(\frac{C_{in}^{(S_j)}}{B_{in}^{(S_j)}}\right)^{C_{in}^{(S_j)}} \left(1 - \frac{C_{in}^{(S_j)}}{B_{in}^{(S_j)}}\right)^{B_{in}^{(S_j)} - C_{in}^{(S_j)}} \left(\frac{C_{out}^{(S_j)}}{B_{out}^{(S_j)}}\right)^{C_{out}^{(S_j)}} \left(1 - \frac{C_{out}^{(S_j)}}{B_{out}^{(S_j)}}\right)^{B_{out}^{(S_j)} - C_{out}^{(S_j)}}$$

which is the numerator in the Bernoulli spatial scan statistic (i.e., Equation 2.1) when

$$\frac{C_{in}^{(S_j)}}{B_{in}^{(S_j)}} > \frac{C_{out}^{(S_j)}}{B_{out}^{(S_j)}}.$$

Since the denominator in that statistic is the same for all subregions, the Bernoulli spatial scan statistic has been derived.

Suppose now that we do not know how many individuals are in the region, and our observations consist only of the individuals who have property X. Then $Data_{in}^{(S_j)}$ consists only of individuals in the *in* region who have property X, and $Data_{out}^{(S_j)}$ consists only of individuals in the *out* region who have property X. We then have

$$P(Data|SUB = S_j) = P(Data_{in}^{(S_j)}|SUB = S_j)P(Data_{out}^{(S_j)}|SUB = S_j) = p^{C_{in}^{(S_j)}} q^{C_{out}^{(S_j)}}$$

Suppose that our estimates of these conditional probabilities are based on the fraction of people who showed the respective behaviors in the subregions. We then have

$$P(Data|SUB = S_j) = \left(\frac{C_{in}^{(S_j)}}{B_{in}^{(S_j)}}\right)^{C_{in}^{(S_j)}} \left(\frac{C_{out}^{(S_j)}}{B_{out}^{(S_j)}}\right)^{C_{out}^{(S_j)}}$$

which is the numerator in the Poisson spatial scan statistic (i.e., Equation 2.2) when

$$\frac{C_{in}^{(S_j)}}{B_{in}^{(S_j)}} > \frac{C_{out}^{(S_j)}}{B_{out}^{(S_j)}}.$$

3.2.3 A Spatial Extension of PC (PCS)

In this section PC is extended to be a spatial model within the BayesNet-S class of models.

3.2.3.1 The Model Figure 3.7 shows the BayesNet-S model obtained by extending PC (Figure 3.3) to a spatial model.

3.2.3.2 The Inference Algorithm Define the following variables:

$Data_{in}^{(S_j)}$: The data on individuals in S_j.
$Data_{out}^{(S_j)}$: The data on individuals in $G - S_j$.

Then from the Bayesian Network in Figure 3.7, we see that

$$P(Data|SUB = none) = P(Data|E = no).$$

To compute this value, we simply use Equation 3.1.

For each subregion S_j and outbreak disease d we have the following:

$$P(Data|SUB = S_j, O = d)$$

$$= P(Data_{in}^{(S_j)}|SUB = S_j, O = d)P(Data_{out}^{(S_j)}|SUB = S_j, O = d) \tag{3.9}$$

The value of

$$P(Data_{in}^{(S_j)}|SUB = S_j, O = d)$$

can be obtained using Equation 3.3, while restricting our data to $Data_{in}^{(S_j)}$. The value of

$$P(Data_{out}^{(S_j)}|SUB = S_j, O = d)$$

can be obtained using Equation 3.1, while restricting our data to $Data_{out}^{(S_j)}$.

We then have that

$$\begin{aligned} P(Data|SUB = S_j) &= \sum_{d \neq none} P(Data|SUB = S_j, O = d)P(O = d|SUB = S_j) \\ &= \sum_{d \neq none} P(Data|SUB = S_j, O = d)P(O = d|E = yes). \end{aligned}$$

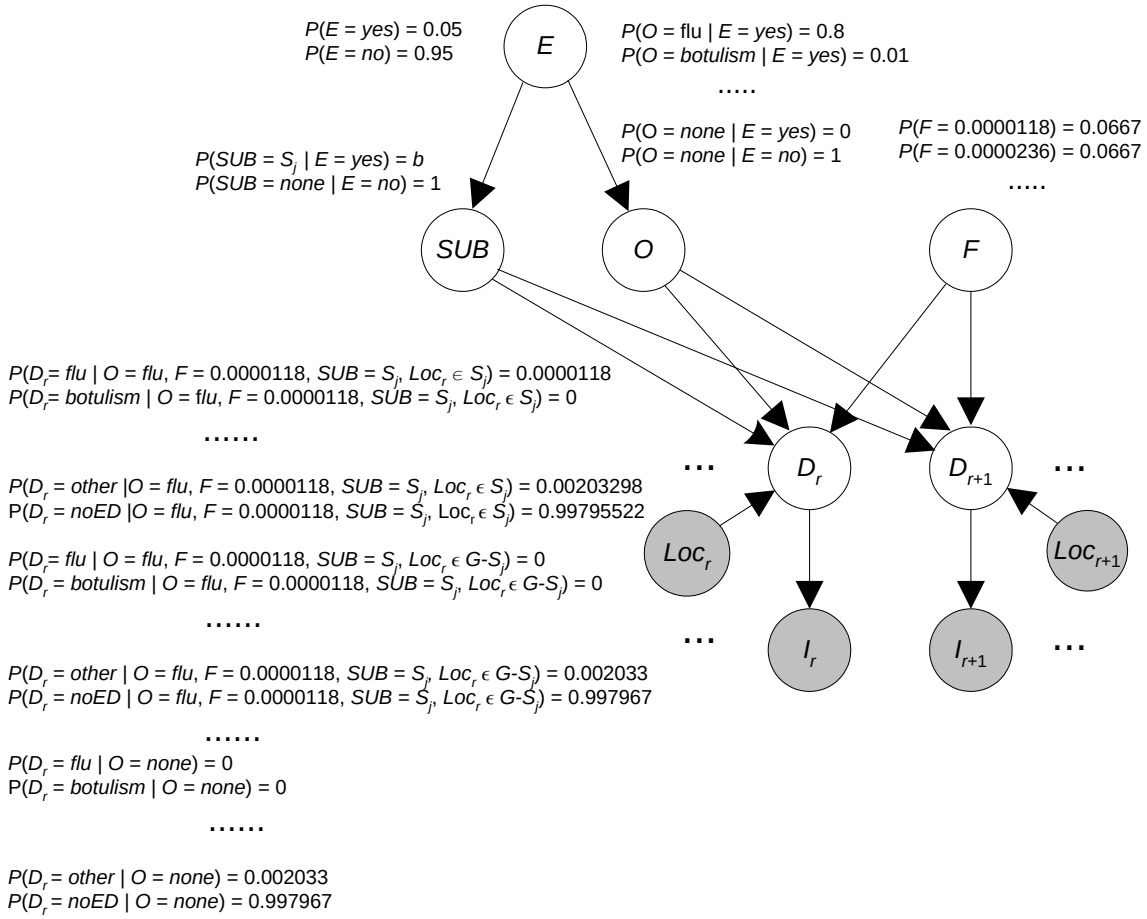


Figure 3.7: The BayesNet-S model obtained by extending PC to a spatial model. The conditional probability distributions for node I_r are the same as those in the Bayesian network for PC, which appears in Figure 3.3.

$P(O = d|E = yes)$ is equal to $P(O = d|SUB = S_j)$ because $P(E = yes|SUB = S_j) = 1$, and E d-separates SUB from O.

We then use Bayes' Theorem to compute

$$P(SUB = S_j|Data) = \frac{P(Data|SUB = S_j)P(SUB = S_j)}{\sum_i P(Data|SUB = S_i)P(SUB = S_i)}.$$

The probability that there is an event in some subregion is as follows:

$$\sum_{S_i \neq none} P(SUB = S_i|Data).$$

In this application, we also want to know the probability of each type of event (outbreak). We have that

$$\begin{aligned} P(Data|O = d) &= \sum_{S_i \neq none} P(Data|SUB = S_i, O = d)P(SUB = S_i|O = d) \\ . &= \sum_{S_i \neq none} P(Data|SUB = S_i, O = d)P(SUB = S_i) \\ &= \sum_{S_i \neq none} P(Data|SUB = S_i, O = d) \times b, \end{aligned}$$

where $1/b$ is the number of subregions. The terms in the expression on the right above have already been computed in Equation 3.9. Furthermore,

$$P(Data|O = none) = P(Data|SUB = none),$$

which has already been computed.

Using Bayes' Theorem,

$$P(O = d|Data) = \frac{P(Data|O = d)P(O = d)}{\sum_c P(Data|O = c)P(O = c)},$$

where c is taken over all its possible values including "none".

Finally,

$$P(E = yes|Data) = \sum_{d \neq none} P(O = d|Data).$$

3.2.3.3 A Time Complexity Analysis of the Inference Algorithm

Define the following variables:

N_C: Number of chief complaints.

N_F: Number of values of F.

N_D: Number of outbreak diseases.

N_S: Number of subregions.

N_{ED}: Number of individuals who visited the ED.

Recall from Section 3.1.3.4 that PC requires $\theta(N_C \times N_F \times N_D)$ time. Since we need to do the computations in PC for every subregion S, the running time of this algorithm is

$$\theta(N_{ED} + N_S \times N_C \times N_F \times N_D).$$

The experiments described in Chapter 4 will consider a monitored region using an $n \times n$ grid similar to the one shown in Figure 2.4. Each grid element is called a cell, and every subset of cells represents a subregion. However, only subregions that are rectangles are investigated. The number of rectangular subregions is

$$\frac{n^2(n+1)^2}{4}.$$

So the running time in terms of the grid size is

$$\theta(N_{ED} + n^4 \times N_C \times N_F \times N_D).$$

3.3 THE BAYESNET-T CLASS OF TEMPORAL EVENT SURVEILLANCE MODELS

This section presents a description of the high-level Bayesian network architecture representing the BayesNet-T class of temporal event surveillance models and then gives a concrete example.

3.3.1 The High-Level Bayesian Network Architecture

We start with the high-level Bayesian network architecture in Figure 3.1. Then two additional random variables, Y and F, are added to the set of attribute variables. These variables are defined as follows:

F: severity or extent of the outbreak if there is an ongoing outbreak.

Y: number of days into the outbreak, if there is an ongoing outbreak.

The specific nature of the variable F depends on the particular application. As to the intermediate and observable variables, there are a set of these variables for today (day 0) and for each day preceding today (day i denotes i days prior to the current day). Their probability distributions are conditional on the values of F, Y, and the day i. The nature of this dependence also depends on the application. The data on day i is denoted $Data(i)$. A high-level Bayesian network architecture representing this class of models appears in Figure 3.8. Any model in this class is called a **Bayesian Network Temporal (BayesNet-T) model.**

Note that my temporal model would only be useful in types of event surveillance that are similar to disease outbreak detection in that the severity of some event increases in each time unit (with possibly fluctuations) up to some point. An example of an event in this category would be a storm that may become a category one hurricane or a category two hurricane, etc. An example of an event that would not be in this category would be the event that there is a cluster of a particular type of tree in the forest. The trees in a possible cluster do not change with time, at least not during the time period in which we are doing the investigation. So there is no purpose in doing temporal modeling. Rather we just look for a cluster at one snapshot in time.

3.3.2 A Temporal Extension of PC (PCT)

3.3.2.1 The Model Figure 3.9 shows the BayesNet-T model obtained by extending PC (Figure 3.3) to a BayesNet-T model.

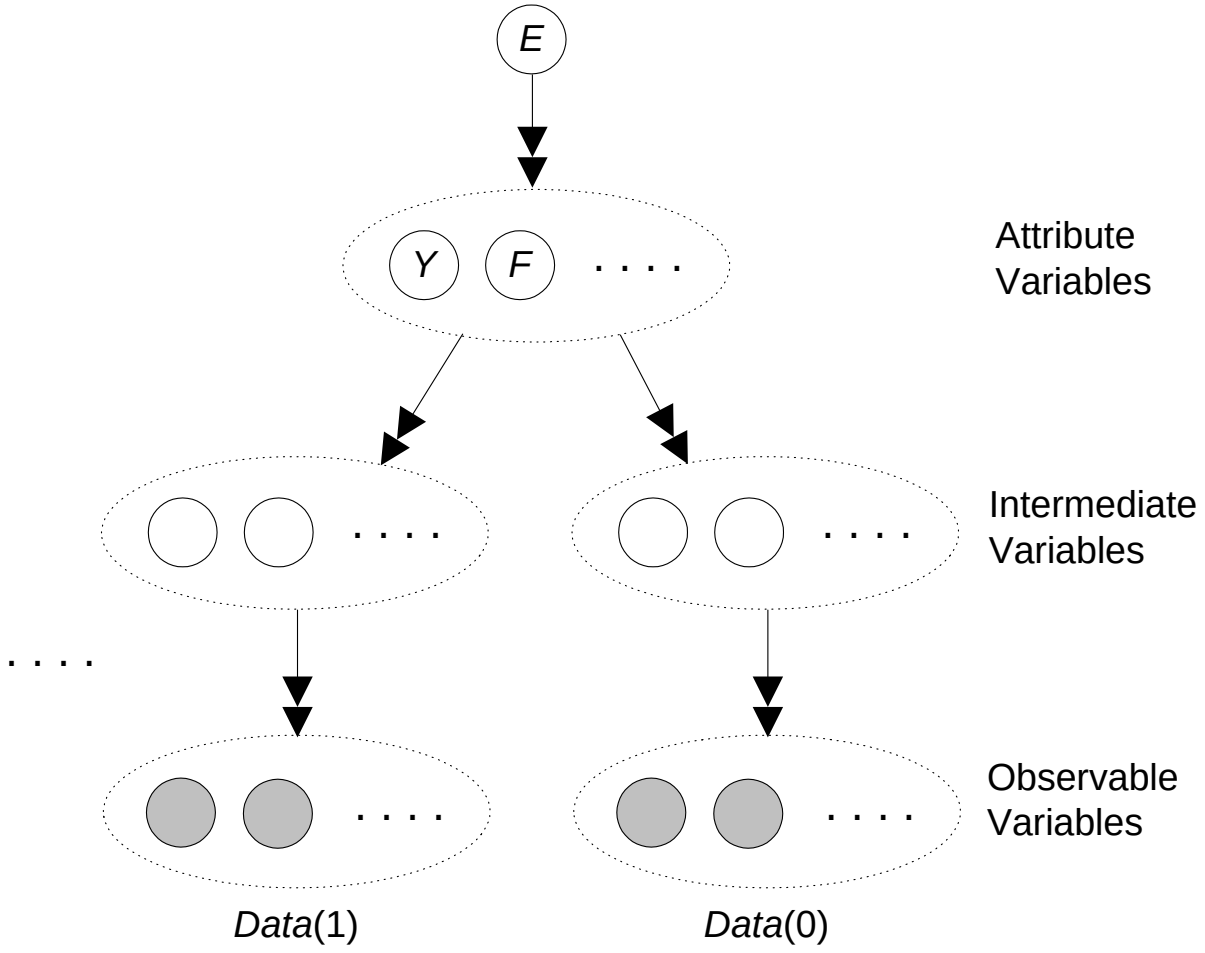


Figure 3.8: The high-level BayesNet-T Bayesian network architecture. The discussion in the caption of Figure 3.1 pertains to this figure. There is always one attribute variable F representing the severity of the outbreak and one attribute variable Y representing the number of days into the outbreak.

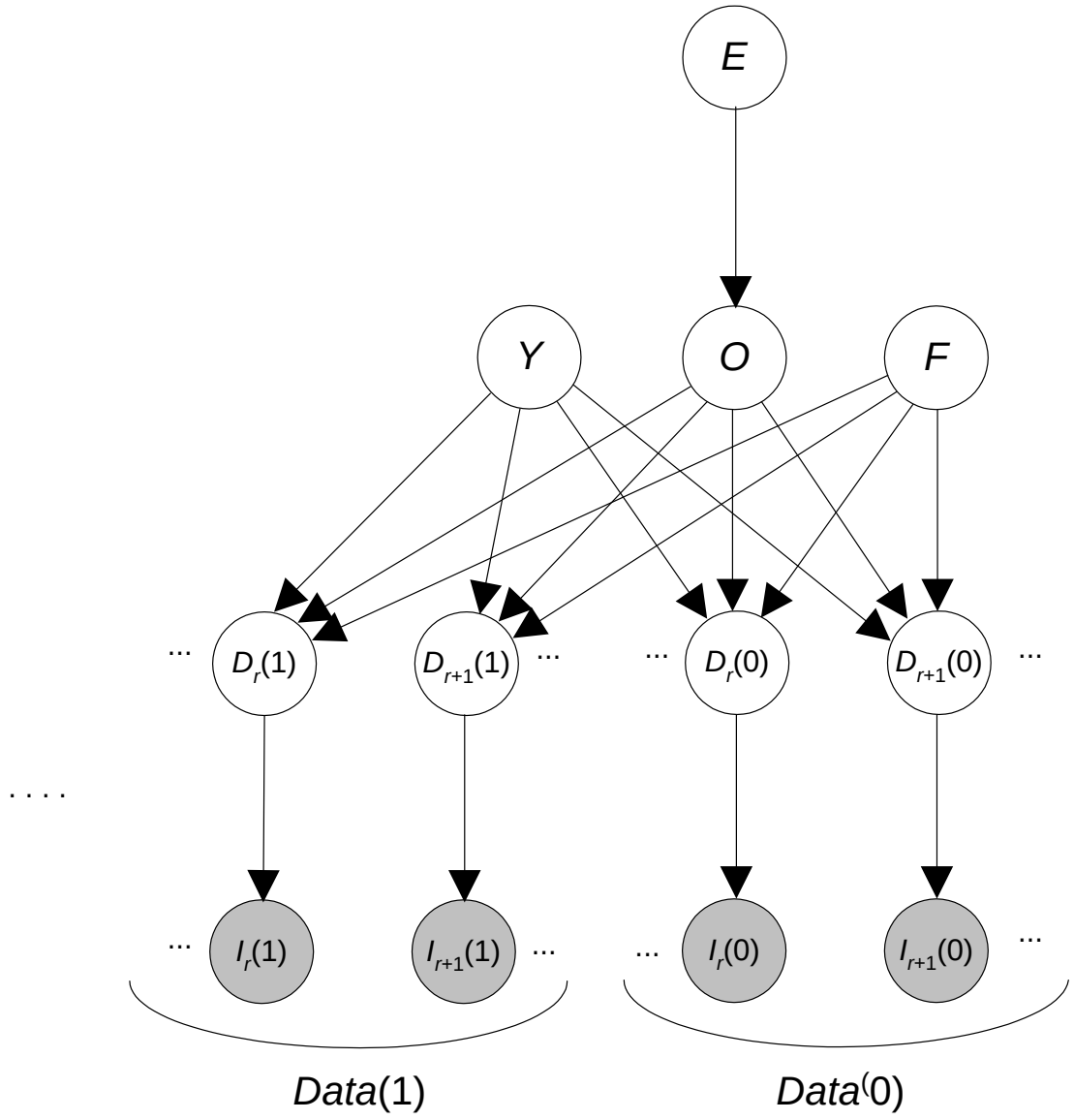


Figure 3.9: The BayesNet-T model obtained by extending PC to a temporal model.

Each day this model bases its outbreak posterior probabilities on the most recent T days[2] (including today) of ED data. I will now describe the nodes in the network.

E: This node represents whether there is an ongoing outbreak. It is the same node as in PC, and its probability distribution is the same as the one in PC.

O: This node represents which outbreak disease is occurring given that there is an outbreak. It is the same node as in PC, and its probability distribution is the same as the one in PC.

F: As in PC, this node represents the probability of an individual both being afflicted with the outbreak disease and showing up today at one of the EDs being monitored, given that an outbreak is ongoing. Its probability distribution is the same as the one in PC.

Y: This node represents the number of days into the outbreak, as of today, if there is an ongoing outbreak. The prior probability over its values is a uniform distribution over $\{1, 2, \ldots T\}$, where T is the maximum time span over which we are modeling an outbreak.

$D_r(i)$: This node represents the ED disease state of the rth individual i days ago, where $i = 0$ represents today. It has the same values as node D_r in PC. Its conditional probability distribution will be discussed shortly.

$I_r(i)$: This node represents the chief complaint of the rth individual i days ago. It has all the same properties as the node I_r in PC. Its conditional probability distributions are the same as those in PC.

The probability distributions of $D_r(i)$ is conditional on O, F, and Y. First, note that the Bayesian network structure in Figure 3.9 entails that, given values of O, F, and Y, the ED disease states (values of $D_r(i)$) and therefore the chief complaints (values of $I_r(i)$) for an individual on different days are independent. For example, conditional on these three variables, if an individual went to the ED yesterday with influenza, it does not change the probability that the individual will go to the ED today with influenza. This assumption allows for a given individual going to the ED two or more times during an outbreak. This

[2] In general, the unit of time need not be a day.

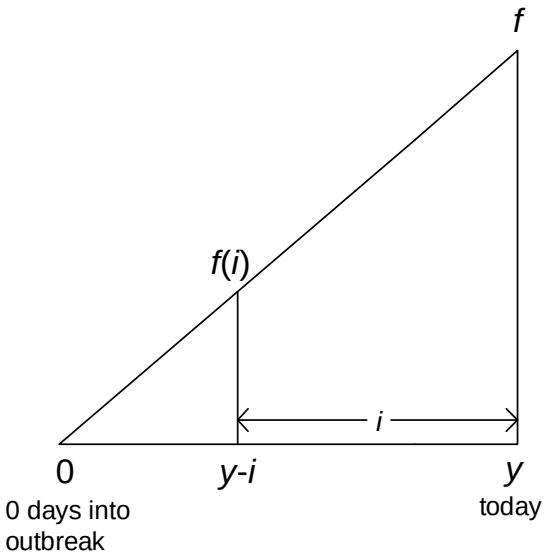


Figure 3.10: Number of days into the outbreak is plotted horizontally, and the prevalence of the outbreak is plotted vertically.

is certainly possible, especially in the case of an outbreak disease with severe symptoms. Further justification for this independence assumption is provided in Section 3.3.2.3 after sufficient notation is introduced.

To develop the conditional probability distribution for node $D_r(i)$, it is useful to first define the following random variable:

$F(i)$: Probability of an individual both being afflicted with the outbreak disease and going to the ED i days ago.

Recall that the value f of F is the probability of an individual both being afflicted with the outbreak disease and going to the ED today, given that an outbreak is ongoing. Early in the outbreak, which is when we hope to detect the outbreak, it is reasonable to assume that the increase in cases can be approximated by a linear increase. Therefore, I assume that the value $f(i)$ of $F(i)$ is related to the values f of F and y of Y as shown in Figure 3.10. This assumption entails that the outbreak extent is at level 0 when we are 0 days into the

outbreak, reaches level f today, and the increase over that period of time is linear. Thus

$$\frac{f(i)}{y-i} = \frac{f}{y},$$

which implies that

$$f(i) = \frac{y-i}{y} f. \tag{3.10}$$

Given Equation 3.10 and the discussion in Section 3.1.3.1 concerning PC, the conditional probability distributions for $D_r^{(i)}$ are as follows:

$$P(D_r(i) = other | O = none, F = f, Y = y) = p_{other}$$

$$P(D_r(i) = noED | O = none, F = f, Y = y) = 1 - p_{other}$$

$$\begin{aligned} P(D_r(i) = d | O = d, F = f, Y = y) &= \frac{y-i}{y} f \qquad i < y \\ &= 0 \qquad\qquad i \geq y \end{aligned} \tag{3.11}$$

$$P(D_r(i) = c | O = d, F = f, Y = y) = 0 \qquad \text{for } c \neq d$$

$$\begin{aligned} P(D_r(i) = other | O = d, F = f, Y = y) &= p_{other} \left(1 - \frac{y-i}{y} f\right) \qquad i < y \\ &= p_{other} \qquad\qquad\qquad\qquad i \geq y \end{aligned} \tag{3.12}$$

$$\begin{aligned} P(D_r(i) = noED | O = d, F = f, Y = y) &= (1 - p_{other}) \left(1 - \frac{y-i}{y} f\right) \qquad i < y \\ &= 1 - p_{other} \qquad\qquad\qquad\qquad i \geq y. \end{aligned} \tag{3.13}$$

Let us discuss the boundary condition in Equation 3.11. Recall that Y is uniformly distributed between 1 and T. It is assumed that we must be at least 1 day into the outbreak for the individual to contract the disease and arrive with it in the ED. The value of $D_r(i)$ is the rth individual's ED disease state i days ago. If $i \geq y$, it is the individual's disease state before we are into the outbreak. For example, if $y = 1$, we are one day into they outbreak today, and so if $i \geq 1$, then i days ago we had not progressed into the outbreak yet. Therefore, in that case the probability of the rth individual having the outbreak disease is 0 as Equation 3.11 entails. A similar discussion pertains to the boundary conditions in Equations 3.12 and 3.13.

3.3.2.2 The Inference Algorithm On each day i, we know the value of $I_r(i)$ for each individual r in the population. $Data(i)$ is the set of these values for day i, and $Data$ is the set of all these values.

Since the data items are conditionally independent given that $O = none$, we have that

$$P(Data|E = no) = P(Data|O = none) = \prod_{i=0}^{T-1} P(Data(i)|O = none). \tag{3.14}$$

Note that the product goes from 0 to $T - 1$, which means we look at T days of data. The terms in the product on the right in Equation 3.14 are given by

$$P(Data(i)|O = none) = \prod_k \left(P(I_r(i) = m_k|O = none)^{C_k(i)}\right.,$$

where $C_k(i)$ is the number of individuals i days ago with chief complaint m_k.

The value of $P(I_r(i) = m_k|O = none)$ is the same as the value of $P(I_r = m_k|O = none)$, which is given by Equation 3.2. So

$$P(I_r(i) = m_k|O = none)$$

$$= P(I_r(i) = m_k|D_r(i) = other) \times p_{other} + P(I_r(i) = m_k|D_r(i) = noED) \times (1 - p_{other}).$$

Next, we have that

$$P(Data|OD = d) = \sum_{f,y} \prod_{i=0}^{T-1} P(Data(i)|O = d, F = f, Y = y)P(F = f)P(Y = y). \tag{3.15}$$

The first term in the product on the right in Equation 3.15 is given by

$$P(Data(i)|O = d, F = f, Y = y) = \prod_k \left(P(I_r(i) = m_k|O = d, F = f, Y = y)^{C_k(i)}\right), \tag{3.16}$$

where $C_k(i)$ is the number of individuals i days ago with chief complaint m_k.

The term in the product on the right in Equation 3.16 above is computed as follows:

$$
\begin{aligned}
& P(I_r(i) = m_k | O = d, F = f, Y = y) \\
&= \sum_c P(I_r(i) = m_k | D_r(i) = c) P(D_r(i) = c | O = d, F = f, Y = y) \\
&= P(I_r(i) = m_k | D_r(i) = d) P(D_r\,(i) = d | O = d, F = f, Y = y) + \\
&\quad P(I_r(i) = m_k | D_r(i) = other) P(D_r(i) = other | O = d, F = f, Y = y) + \\
&\quad P(I_r(i) = m_k | D_r(i) = noED) P(D_r(i) = noED | O = d, F = f, Y = y).
\end{aligned} \tag{3.17}
$$

The conditional probabilities of values of $D_r^{(i)}$ in the previous expression are computed using Equations 3.11, 3.12, and 3.13.

Using Bayes' Theorem, we have that

$$P(O = d | Data) = \frac{P(Data | O = d) P(O = d)}{\sum_c P(Data | O = c) P(O = c)}.$$

Finally,

$$P(E = yes | Data) = \sum_{d \neq none} P(O = d | Data).$$

3.3.2.3 Justification for the Independence Assumption Recall that the model assumes that, given values of O, F, and Y, the chief complaints of an individual on different days are independent. I will now discuss the justification for this assumption, which serves as an approximation.

First, I assume that the number of days of data T (time window) is fairly small. In my experiments I used $T = 5$. For the sake of simplicity, in the discussion that follows, I use $T = 3$. Also, for the sake of brevity, in what follows I denote the event

$$O = d, F = f, Y = y$$

by e.

Next, note that $P(I_r = NoED | e) \approx 1$ for any set of values of $O = d, F = f$, and $Y = y$. This can been seem by looking at Figure 3.3. We see from that figure that, for example,

$$P(D_r = noED | O = flu, F = 0.0000118) = .9979552 \tag{3.18}$$

$$P(D_r = noED | O = none, F = 0.0000118) = .997967 \tag{3.19}$$

$$P(I_r = noED | D_r = noED) = 1. \tag{3.20}$$

Equations 3.18 and 3.19 are exemplary of what is true in general. Namely, regardless of the value of O or F (and of Y in the temporal model), an individual will most probably not go to the ED. We see from Equation 3.20 that if an individual does not go to the ED, then the value of I_r is noED.

Now I will discuss the justification for the independence assumption separately for three types of individuals.

1. Individual r does not visit the ED during the time window.

   My assumption relevant to these individuals is that if an individual does not go to the ED on one or more days in a row, it is still most probable that the individual will not go to the ED the following day. Let $I_r(k)$ be the individual's chief complaint on the kth day. Due to the chain rule, we then have that (recall that for simplicity we assumed $T = 3$)

$$\begin{aligned} & P(I_r(0) = noED, I_r(1) = noED, I_r(2) = noED | e) \\ = \; & P(I_r(0) = noED | I_r(1) = noED, I_r(2) = noED, e) \times \\ & P(I_r(1) = noED | I_r(2) = noED, e) \times P(I_r(2) = noED | e) \\ \approx \; & 1 \\ \approx \; & P(I_r(0) = noED | e) \times P(I_r(1) = noED | e) \times P(I_r(2) = noED | e). \end{aligned} \tag{3.21}$$

   The approximations hold because all terms in the first product are assumed to be near 1, all terms in the second product are near 1, and we have assumed the time window is small. Now

$$P(I_r(0) = noED, I_r(1) = noED, I_r(2) = noED | e)$$

   is the actual conditional probability of the data concerning individual r, and

$$P(I_r(0) = noED | e) \times P(I_r(1) = noED | e) \times P(I_r(2) = noED | e)$$

   is the value used by my model. So my assumption implies that the value used by my model is a close approximation.

The assumption made here is wrong only if not going to the ED several days in a row somehow made it probable an individual would go to the ED the following day, which does not seem reasonable. So the assumption concerning these individuals is cogent, and this assumption concerns most of the individuals in the population since most individuals do not visit the ED in a short time window (unless there was a very severe outbreak in which case computer-assisted outbreak detection would probably not be needed).

It is believed that no outbreak occurred in Allegheny County during the calendar year 2004. Using ED data from that county during that calendar year, I evaluated the accuracy of the assumption concerning these individuals as follows. I determined whether these data indicated that Equation 3.21 held for $T = 5$ and

$$e = (O = none, F = f, Y = y).$$

To accomplish this, for each day j in the 365 day period I counted the number N_j of individuals who did not visit the ED. I then estimated that

$$P(I_r(k) = noED|e) = \frac{\sum_{j=1}^{365} \left(\frac{N_j}{N}\right)}{365} = \frac{\sum_{j=1}^{365} N_j}{365 \times N}, \tag{3.22}$$

where N is the number of individuals in the population and $I_r(k)$ is individual r's chief complaint on the kth day. This value is an estimate of the probability of an individual not visiting the ED on any given day. My estimate of the product on the right in Equation 3.21 when $T = 5$ is then

$$\left(\frac{\sum_{j=1}^{365} N_j}{365 \times N}\right)^5.$$

To estimate the value on the left in Equation 3.21 when $T = 5$, I looked at the 361 five-day windows in the time period. That is, first I looked at January 1, 2004 - January 5, 2004, next I looked at January 2, 2004 - January 6, 2004, and so on. For each window i, I counted the number N_i of individuals who did not visit the ED during the entire window. My estimate of the value on the left in Equation 3.21 is then

$$\frac{\sum_{i=1}^{361} \left(\frac{N_i}{N}\right)}{361} = \frac{\sum_{i=1}^{361} N_i}{361 \times N}.$$

My results were as follows:

$$\left(\frac{\sum_{j=1}^{365} N_j}{365 \times N}\right)^5 = 0.99006$$

$$\frac{\sum_{i=1}^{361} N_i}{361 \times N} = 0.99032.$$

These numbers are equal to the third decimal place.

2. Individual r visits the ED once during the the time window.

My assumption relevant to these individuals is that if an individual goes to the ED on day 2, the probability of not going to the ED on day 1 remains very high. Furthermore, if the individual goes to the ED on day 2, and does not go to the ED on day 1, the probability of not going to the ED on day 0 remains high. In general, the assumption is that if an individual goes to the ED on day i and does not go the ED on days $j+1, ..., i-1$, then the probability of not going to the ED on day j remains high, where $j < i$ and j and i are both in the window. This assumption seems reasonable. If an individual goes to the ED one day, one might argue that it would increase the probability of going to the ED a second day because the individual is sick. Or one might argue that it would decrease the probability of going to the ED another day because the individual has already been to ED. Regardless of which of these might be correct, it does not seem like the probability would change much. My assumption would only be incorrect if an ED visit on one day made the probability of an ED visit another day substantially different. So this assumption is reasonable, but perhaps not at compelling as the assumption for Type 1 individuals which was discussed above.

Without loss of generality, assume that the individual's sole ED visit is on day 2 and that the chief complaint is m_k. Given the assumption above, due to the chain rule we then have that

$$\begin{aligned}
&P(I_r(0) = noED, I_r(1) = noED, I_r(2) = m_k|e) \\
&\quad = P(I_r(0) = noED|I_r(1) = noED, I_r(2) = m_k, e) \times P(I_r(1)|I_r(2) = m_k, e) \times \\
&\qquad P(I_r(2) = m_k|e) \\
&\quad \approx P(I_r(2) = m_k|e) \\
&\quad \approx P(I_r(0) = noED|e) \times P(I_r(1) = noED|e) \times P(I_r(2) = m_k|e). \qquad (3.23)
\end{aligned}$$

Now

$$P(I_r(0) = noED, I_r(1) = noED, I_r(2) = m_k|e)$$

is the actual conditional probability of the data concerning individual r, and

$$P(I_r(0) = noED|e) \times P(I_r(1) = noED|e) \times P(I_r(2) = m_k|e)$$

is the value used by my model. So my assumption implies that the value used by my model is a close approximation.

Again using ED data from Allegheny County during the calendar year 2004, I evaluated the accuracy of the assumption concerning these individuals as follows. I determined whether these data indicated that Equation 3.23 held for $T = 5$,

$$e = (O = none, F = f, Y = y),$$

and one particular common chief complaint, namely "cough." To accomplish this, I first estimated that

$$P(I_r(2) = cough|e) = \frac{\sum_{j=1}^{365} \left(\frac{M_j}{N}\right)}{365} = \frac{\sum_{j=1}^{365} M_j}{365 \times N}, \tag{3.24}$$

where M_j is the number of individuals who presented in the ED with a cough on the jth day of the period. For each five-day window i, I counted the number K_i of individuals who did not visit the ED on four of those days and who visited the ED with a cough on the other day. I then computed

$$\frac{\sum_{i=1}^{361} \left(\frac{K_i}{N}\right)}{361} = \frac{\sum_{i=1}^{361} K_i}{361 \times N},$$

and compared the result to

$$5 \left(\frac{\sum_{j=1}^{365} N_j}{365 \times N}\right)^4 \left(\frac{\sum_{j=1}^{365} M_j}{365 \times N}\right).$$

In this last expression, the term on the left is from Equation 3.22, and the factor 5 is present because the ED visit could occur on any of the five days in the window.

My results were as follows:

$$\frac{\sum_{i=1}^{361} K_i}{361 \times N} = 0.0002508$$

$$5\left(\frac{\sum_{j=1}^{365} N_j}{365 \times N}\right)^4 \left(\frac{\sum_{j=1}^{365} M_j}{365 \times N}\right) = 0.000\,2540\,.$$

These number are equal to the fifth decimal place.

3. Individual r visits the ED more than once during the time window.

   Assuming conditional independence means that the naive Bayes assumption is being made. Although in many cases this assumption is not literally true, it often has been shown to perform well in practice on classification tasks (Sun and Shenoy, 2006).

   Again using ED data from Allegheny County during the calendar year 2004, I evaluated the accuracy of the assumption concerning these individuals as follows. In the same way as done for Type 2 individuals, I investigated the assumption for one particular common chief complaint, namely "cough". For each five-day window i, I let L_i be the number of individuals who did not visit the ED on three of those days and who visited the ED with a cough on the other two days. I then computed

$$\frac{\sum_{i=1}^{361} \left(\frac{L_i}{N}\right)}{361} = \frac{\sum_{i=1}^{361} L_i}{361 \times N},$$

   and compared the result to

$$10\left(\frac{\sum_{j=1}^{365} N_j}{365 \times N}\right)^3 \left(\frac{\sum_{j=1}^{365} M_j}{365 \times N}\right)^2 .$$

   In this last expression, the term on the left is from Equation 3.22, the term on the right is from Equation 3.24, and the factor 10, which is the value of the binomial coefficient $\binom{5}{2}$, is present because the two ED visits could occur on any of the two days in the five-day window.

   My results were as follows:

$$\frac{\sum_{i=1}^{361} L_i}{361 \times N} = 2.027 \times 10^{-6}$$

$$10\left(\frac{\sum_{j=1}^{365} N_j}{365 \times N}\right)^3 \left(\frac{\sum_{j=1}^{365} M_j}{365 \times N}\right)^2 = 2.6 \times 10^{-8}.$$

   The value obtained using the independence assumption is two orders of magnitude smaller than the other value. The independence assumption substantially underestimates the probability of the data. However, it seems plausible that it should underestimate the

probability of the data in a similar way for all hypotheses because each of them assumes the same type of independence; thus, the net effect of the underestimation would likely be attenuated.

Since it seems that the assumption for Type 3 individuals may not hold, I investigated what fraction of individuals fall into this category. My investigation proceeded as follows. Again, I analyzed ED data from Allegheny County during the calendar year 2004. For each five-day window i and for $0 \leq j \leq 5$, I determined $N_i^{(j)}$, which is the total number of individuals visiting the ED j times during the window. I then estimated for $0 \leq j \leq 5$ that

$$P(Visits = j) = \frac{\sum_{i=1}^{361} \frac{N_i^{(j)}}{N}}{361} = \frac{\sum_{i=1}^{361} N_i^{(j)}}{361 \times N},$$

where N is the number of individuals in the population, and $P(Visits = j)$ is an estimate of the probability of an individual visiting the ED j times during a five-day window. The resultant distribution is as follows:

$$\begin{aligned}
P(Visits &= 0) = 0.990320292 \\
P(Visits &= 1) = 0.009359424 \\
P(Visits &= 2) = 0.000298369 \\
P(Visits &= 3) = 0.000018714 \\
P(Visits &= 4) = 0.000002512 \\
P(Visits &= 5) = 0.000000686.
\end{aligned}$$

The probability of an individual not going to the ED is close to 1. The probability of an individual going to the ED one time is two orders of magnitude smaller than the probability of not going to the ED. The probability decreases exponentially thereafter for higher numbers of visits. The probability of an individual going to the ED two or more times is equal to

$$0.000298369 + 0.000018714 + 0.000002512 + 0.000000686 = 0.000320281,$$

which is about 29 times less than the probability of an individual going exactly once to the ED. I conclude that a very small percentage of individuals are Type 3.

The independence assumption being made holds well in those cases that are most likely to happen (no visits or one visit per period). Those cases in which it is not well satisfied are relatively uncommon (more than one visit per period). Even in those cases, we might expect that the net effect of the independence assumption would be attenuated because all the hypotheses assume the same type of independence; this is an issue for further study.

In another domain this independence assumption may not hold. For example, suppose we are monitoring computers on a network for the emergence of a virus. If a particular computer exhibited a manifestation of a virus on one day, the probability of such a manifestation on the next day would become greater (because the probability of a virus spreading has increased). So my independence assumption may not hold well.

3.3.2.4 A Time Complexity Analysis of the Inference Algorithm Define the following variables:

N_C: Number of chief complaints.

N_F: Number of values of F.

N_D: Number of outbreak diseases.

N_{ED}: Number of individuals who visited the ED in the past T days (including today).

T: Number of days of data investigated by the model. This variable was previously defined. Note that this is also the number of different values of random variable Y.

For a given value of d, it is necessary to compute $N_F \times T^2$ terms using Equality 3.15. Each of these computations requires that we compute N_C terms using Equality 3.16. Each term in Equality 3.16 can be computed in constant time using Equality 3.17. Since there are N_D outbreak diseases, we conclude that the time complexity of the algorithm is

$$\theta(N_{ED} + N_C \times N_F \times N_D \times T^2).$$

As in Section 3.1.3.4, we have included the time to pre-process the data.

3.3.2.5 A Comparison to Other Methods As mentioned in Section 2.3.3.2, a temporal version of the spatial scan statistic appears in (Kulldorff, 2001) and (Kulldorff et al., 2005). It is like the basic version except that instead of a circular window in two dimensions, the space-time scan statistic searches over cylindrical windows in three dimensions. This method does not model the phenomenon that counts ordinarily increase during a disease outbreak. It was further mentioned in Section 2.3.3.2 that (Neill et al., 2005b) developed a temporal version of the spatial scan statistic that does model increasing counts. This version looks at counts from previous days, and, in the alternative hypothesis, uses a different parameter p for each day, where the values of these parameters are assumed to be monotonically increasing. My extension of PC to PCT is similar to this latter frequentist model because I assumed that the prevalence of the outbreak (value of f) is increasing linearly. However, in general a (BayesNet-T) model could assume a non-linear increase, a decrease, or even a constant value**.** For example, we would assume a constant value in the case of a **persistent cluster**, which is a cluster which starts at some point in time but does not change much after its initial onset. An example, might be a radiation leak in a community. Although it would emerge in space over time, the radiation level at particular location should not change much after the initial onset.

3.4 THE BAYESNET-ST CLASS OF SPATIO-TEMPORAL EVENT SURVEILLANCE MODELS

This section presents a description of the high-level Bayesian network architecture representing the BayesNet-ST class of spatio-temporal event surveillance models and then gives a concrete example.

3.4.1 The High-Level Bayesian Network Architecture

The high-level Bayesian network architecture for spatio-temporal event surveillance is a combination of the high-level Bayesian network architecture for spatial event surveillance in

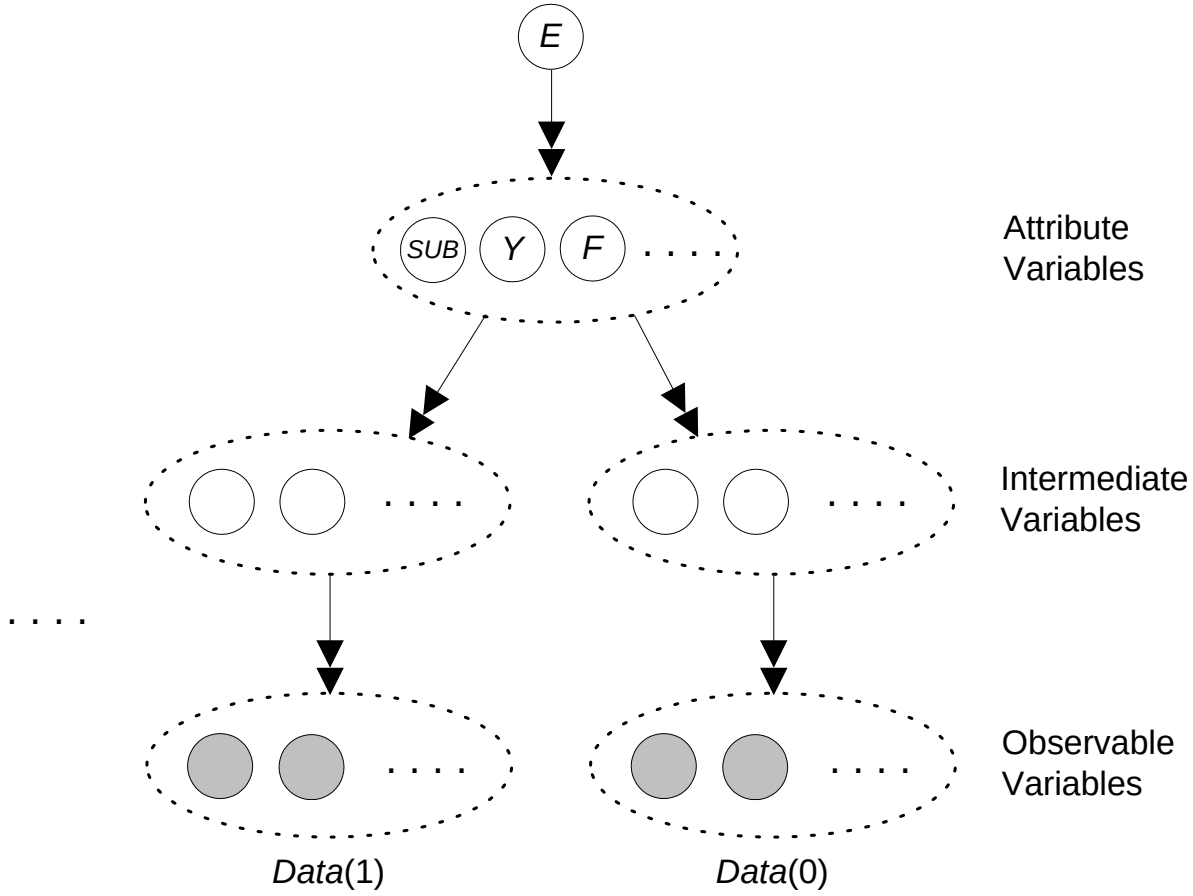


Figure 3.11: The high-level BayesNet-ST Bayesian network architecture. The discussion in the caption of Figure 3.1 pertains to this figure. There is always one attribute variable SUB, whose value is the subregion in which the event is occurring if there is an event, one attribute variable F representing the severity of the outbreak, and one attribute variable Y representing the number of days into the outbreak.

Figure 3.5 and the high-level Bayesian network architecture for temporal event surveillance in Figure 3.8. It appears in Figure 3.11. Any model in the class of models represented by this Bayesian network is called a **Bayesian Network Spatio-Temporal (BayesNet-ST) model**.

3.4.2 A Spatio-Temporal Extension of PC (PCTS)

3.4.2.1 The Model The BayesNet-ST model obtained by extending PC (Figure 3.3) to a spatio-temporal model appears in Figure 3.12. It is a combination of the spatial Bayesian network in Figure 3.7 and the temporal Bayesian network in Figure 3.9.

3.4.2.2 The Inference Algorithm The inference algorithm for this model uses the calculations in the inference algorithm for the temporal model (Section 3.3.2.2) similar to the way the inference algorithm for the spatial model (Section 3.2.3.2) uses the calculations in the inference algorithm for the basic PC model (Section 3.1.3.3).

To compute

$$P(Data|SUB = none) = P(Data|E = no),$$

we simply use Equation 3.14.

For each subregion S and outbreak disease d we have that

$$P(Data|SUB = S, O = d) = P(Data_{in}^{(S)}|O = d)P(Data_{out}^{(S)}|E = no). \tag{3.25}$$

The value of

$$P(Data_{in}^{(S)}|O = d)$$

can be obtained using Equation 3.15, while restricting our data to $Data_{in}^{(S)}$. The value of

$$P(Data_{out}^{(S)}|E = no)$$

can be obtained using Equation 3.14, while restricting our data to $Data_{out}^{(S)}$.

We then have that

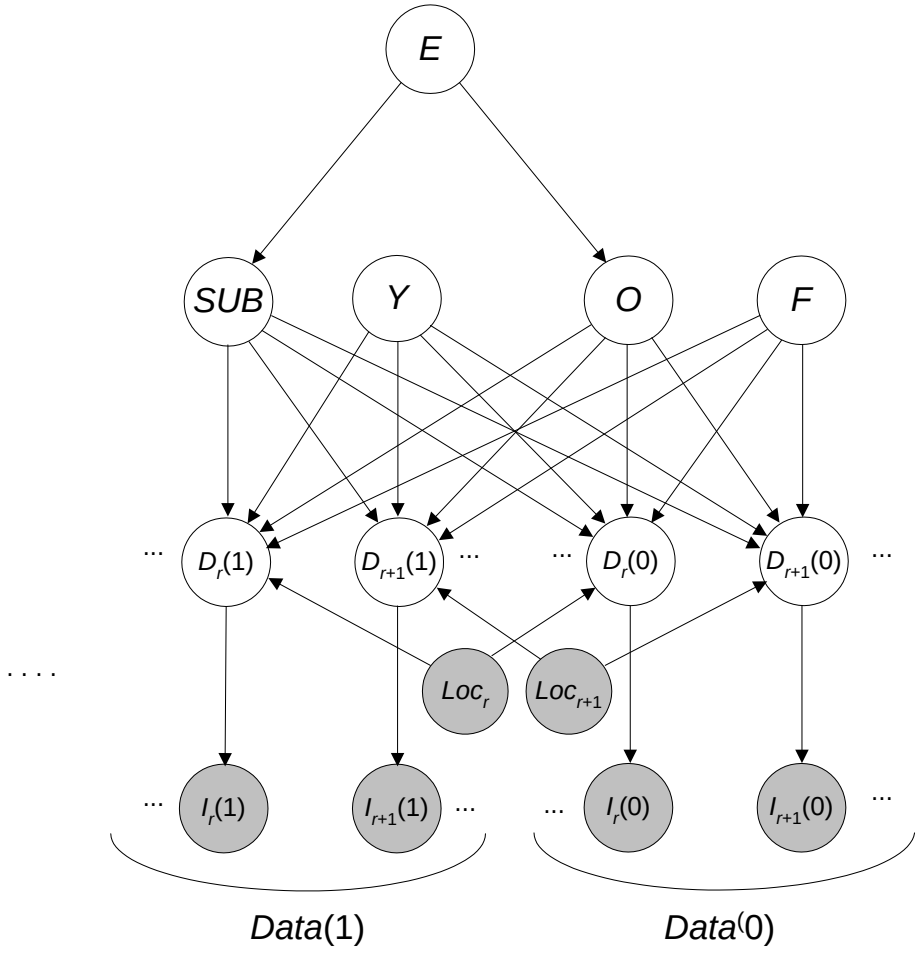


Figure 3.12: The BayesNet-ST model obtained by extending PC to a spatio-temporal model.

$$\begin{aligned} P(Data|SUB = S) &= \sum_{d \neq none} P(Data|SUB = S, O = d)P(O = d|SUB = S) \\ &= \sum_{d \neq none} P(Data|SUB = S, O = d)P(O = d|E = yes). \end{aligned}$$

We then use Bayes' Theorem to compute

$$P(SUB = S|Data) = \frac{P(Data|SUB = S)P(SUB = S)}{\sum_{R} P(Data|SUB = R)P(SUB = R)}.$$

The probability that there is an event in some subregion is equal to

$$\sum_{S \neq none} P(SUB = S|Data).$$

In this application, we also want to know the probability of each type of event (outbreak). We have that

$$P(Data|O = d) = \sum_{S \neq none} P(Data|SUB = S, O = d)P(SUB = S|O = d).$$

The values on the right have already been computed using Equation 3.25. Furthermore,

$$P(Data|O = none) = P(Data|SUB = none),$$

which has already been computed.

Using Bayes' Theorem,

$$P(O = d|Data) = \frac{P(Data|O = d)P(O = d)}{\sum_{c} P(Data|O = c)P(O = c)}.$$

Finally,

$$P(E = yes|Data) = \sum_{d \neq none} P(O = d|Data).$$

3.4.2.3 A Time Complexity Analysis of the Inference Algorithm

Define the following variables:

N_C: Number of chief complaints.

N_F: Number of values of F.

N_D: Number of outbreak diseases.

N_S: Number of subregions.

T: Number of days of data investigated by the model.

N_{ED}: Number of individuals who visited the ED in the past T days (including today).

Combining the analyses in Section 3.2.3.3 and Section 3.3.2.4, we have that the running time of the algorithm is

$$\theta(N_{ED} + N_S \times N_C \times N_F \times N_D \times T^2).$$

Furthermore, in terms of the grid size n, the running time is

$$\theta(N_{ED} + n^4 \times N_C \times N_F \times N_D \times T^2).$$

3.5 ADVANTAGES OF A BAYESNET-ST MODEL

A BayesNet-ST model may show better detection performance (than the spatial scan statistic) because such a model has the following potential advantages:

1. A BayesNet-ST model can readily include multinomial variables. (Jung et al., 2007) developed a spatial scan statistic that handles a multinomial variable, but it requires that the values of the variable be ordinal in nature. That is, we must be able to rank them from lowest to highest. A BayesNet-ST has no such requirement.

2. The spatial scan statistic only investigates whether a cluster of events related to the event of interest is occurring in subregion S. A BayesNet-ST model can use a Bayesian network to model the entire causal mechanism according to which the cluster might occur. For example, in our application to disease outbreak detection, the spatial scan statistic only investigates whether a cluster that is indicative of an outbreak is occurring in a given subregion. A BayesNet-ST model can model the diseases that could be causing the outbreak, the severity of the outbreak, and the relationship of the observed variables to the diseases and the severity.
3. A BayesNet-ST model readily allows more than one observable random variable for each individual. (Kulldorff et al., 2007) developed multivariate versions of the spatial scan statistics. Their method consists of summing the log likelihood ratios over all variables. This amounts to assuming the variables are independent conditional on whether there is a cluster in a given subregion. This assumption is restrictive. An advantage of using a Bayesian network is that a Bayesian network can readily represent the relationships among all variables in the domain, including any known causal relationships. So it is not necessary to assume conditional independence.
4. Related to (3) above, A BayesNet-ST model can include nodes for each individual in the population in the Bayesian network, which means the resultant model is entity-based. As mentioned in Section 2.3.1.2, by using an entity-based model, we can base our analysis on more information than that contained in a summary statistic.

4.0 EXPERIMENTS TESTING THE HYPOTHESES OF THIS THESIS

This chapter first states in detail my hypotheses. Then I discuss the methods used to evaluate the results of the experiments testing the hypotheses. Finally, the experiments and their results are presented.

4.1 HYPOTHESES

As discussed in Section 2.3.1.2, PC is an entity-based, non-spatial, non-temporal outbreak detection system that uses a Bayesian network model. In Section 3.2.3, the Bayesian network in PC was extended to a BayesNet-S model resulting in an entity-based, non-temporal, spatial outbreak detection system called PCS. In Section 3.3.2, the Bayesian network in PC was extended to a BayesNet-T model resulting in an entity-based, temporal, non-spatial outbreak detection system called PCT. In Section 3.4.2 the Bayesian network in PC was extended to a BayesNet-ST model resulting in an entity-based, spatio-temporal outbreak detection system called PCTS. The lattice in Figure 4.1 shows a hierarchical structure for these systems. In that lattice, a parent system X is an enhancement of the systems at X's children.

The hypothesis addressed by this research is that the system at each node in the lattice in Figure 4.1 improves event surveillance relative to the system at each of the node's children. The four specific **hypotheses** are as follows:

1. PCS is an improvement on PC in that it will have a smaller mean time to detection at most false alarm rates. Second, PCS can accurately locate the subregion in which an

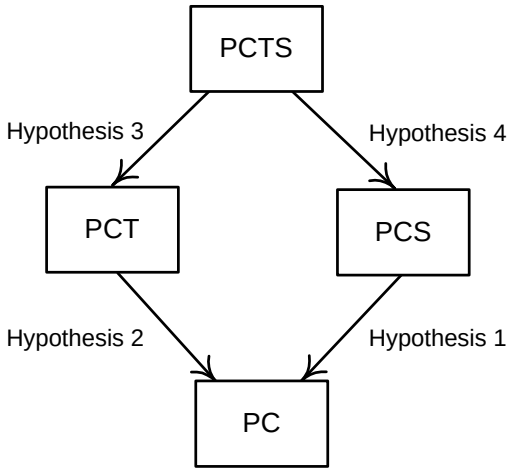


Figure 4.1: It is hypothesized that, as we go up this lattice, event surveillance will improve.

outbreak is occurring when the outbreak is restricted to a subregion of the monitored region.

2. PCT is more stable than PC in that once an outbreak is detected, PCT is better at maintaining the detection signal on future days. Furthermore, PCT can accomplish this without adversely affecting the mean time to detection at each false alarm rate.
3. PCTS is an improvement on PCT in that it will have a smaller mean time to detection at most false alarm rates.
4. PCTS is an improvement on PCS in that once an outbreak is detected, PCTS can better maintain the detection signal on future days. Furthermore, PCTS can accomplish this without adversely affecting the mean time detection at each false alarm rate.

I conjecture that the above hypotheses are true for the following reasons: 1) PC and PCT look for an outbreak only in an entire region, and therefore could easily overlook that the number of new cases is high in a small subregion. PCS and PCTS look at the entire region and at subregions. So, when an outbreak starts in a small subregion, they should usually exhibit better early detection capability. Furthermore, their ability to detect an outbreak occurring in a subregion should not compromise their ability to detect an outbreak occurring

in the entire region. The reason is that PCS and PCTS do not look for outbreaks by checking whether a count in a subregion is greater than the count outside the subregion. Rather, like PC and PCT, they base detection on whether there is a deviation from what is usual. So PCS and PCTS should exhibit early detection capability similar to that of PC and PCT, respectively, even when there are no significant clusters in the region. 2) PC and PCS only look at each day's data. During an outbreak, the number of new outbreak cases each day shows an overall increase as we proceed into the outbreak, but the daily fluctuations in the number of cases can be dramatic. Therefore, a system that looks only at each day's data might signal an outbreak one day because the number of new outbreak cases that day is high, and not signal one the next day because the number has dropped back down. However systems like PCT and PCTS, which looks at data from preceding days, would see that the number of new cases was high yesterday, perhaps 3 days ago, etc., and thus maintain the signal that there is an outbreak on a day when the number had dropped back down.

4.2 EVALUATION METHODOLOGY

A number of outbreak detection systems were compared in the experiments discussed in this chapter and in Chapter 5. All the experiments involved simulated outbreaks that use semi-synthetic data. Presently, the properties of the simulations that are common to all the experiments are discussed, and the methodologies used to evaluate the results of the experiments are presented.

4.2.1 The Simulations

Allegheny County, Pennsylvania, which covers 730 square miles, was modeled using a 16×16 grid similar to the one shown in Figure 2.4. Each grid element is one cell. A zip code was considered entirely within a cell if the zip code's centroid was in the cell. The actual grid, along with a rectangular subregion, appears in Figure 4.2. Outbreaks were simulated in rectangular subregions of that county. In all the experiments, influenza and *Cryptosporidium*

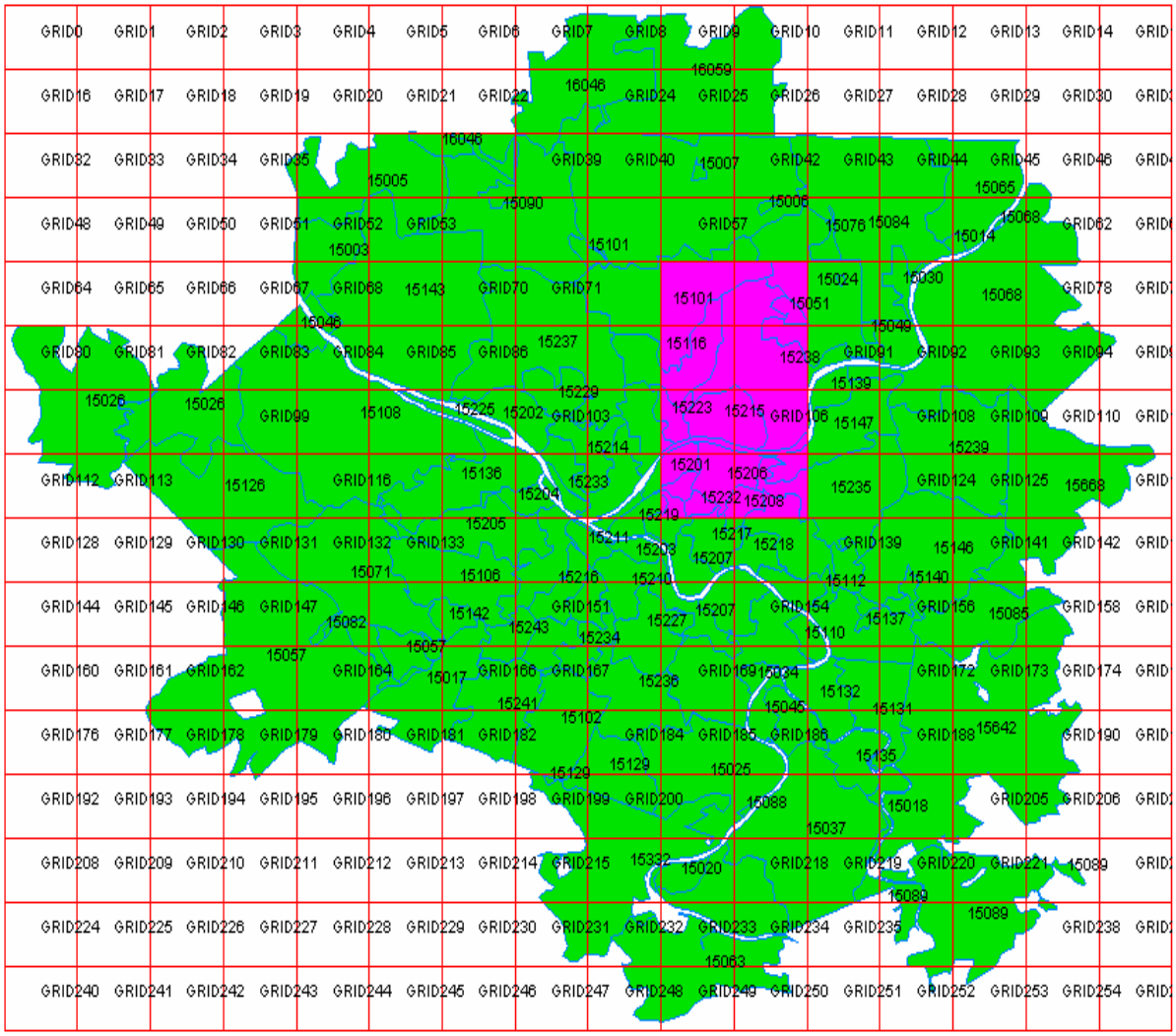


Figure 4.2: Allegheny County is covered with a 16×16 rectangular grid. Each grid element is one cell. A zip code was considered entirely within a cell if the zip code's centroid was in the cell.

outbreaks were simulated because outbreaks of these types have been well studied (Stirling et al., 2001), (Jiang, 2006), (Cooper et al., 2007). The observed data for both types of outbreaks consisted of chief complaints presented by patients in the ED.

Each outbreak was simulated by injecting new ED visits into a portion of a one-year background period when it was assumed no outbreak was occurring. The period chosen was the entire calendar year 2004. The data for each outbreak consisted of the new injected ED visits plus the ED data in the background period. So the data was semi-synthetic because the background data is real and the overlaid outbreak data is synthetic.

Even though some of the systems (e.g., PC and PCS) do not analyze patterns of evidence over time, the purpose of all the systems is to detect outbreaks that emerge in time. So emergence in time was simulated. It was assumed that an outbreak started on day 0 of the outbreak, and that manifestations of outbreak patients were first observed on day 1 of the outbreak, which is said to be 1 day into the outbreak. The number of outbreak patients with observed manifestations then increased each day after that, except that in some of the outbreaks daily fluctuation was modeled.

4.2.2 AMOC Curves

AMOC curves (Fawcett and Provost, 1999) were used to evaluate the ability of the systems to detect the outbreaks. In such curves, the annual number of false alarms is plotted on the x-axis and the mean days to detection is plotted on the y-axis.

Points for the AMOC curves can be obtained as follows. To produce an AMOC curve for a given system and a given set of outbreaks, first the system is run every day during the one-year background period, and the value of $P(E = yes|Data)$ is computed for each of these days (note that $Data$ depends on the day). Then these posterior probabilities are ordered in decreasing order. Let $threshold_0$, $threshold_1, \ldots,$ $threshold_{364}$ be that ordered list, where duplicate values each occupy their own spot in the list. For example, our list may

be as follows:

$$\begin{aligned}
&0.95\\
&0.9\\
&0.8\\
&0.8\\
&0.72\\
&\vdots\\
&0.01.
\end{aligned}$$

For a given outbreak B, let $Data_i^{(B)}$ be the data obtained on the ith day of the outbreak. For a given false alarm rate r, let i_r be the smallest value of i such that

$$P(E = yes|Data_i^{(B)}) > threshold_r. \tag{4.1}$$

For outbreak B, the days to detection is then equal to i_r at an annual false alarm rate of r. Note that, when there are duplicate values in the list, we only obtain a point on the AMOC curve for the first occurrence of the value.

Example 4.1. *Suppose we have the ordered list above. Then* $threshold_2 = 0.8$. *Suppose*

$$P(E = yes|Data_1^{(B)}) = 0.11$$

$$P(E = yes|Data_2^{(B)}) = 0.53$$

$$P(E = yes|Data_3^{(B)}) = 0.73$$

$$P(E = yes|Data_4^{(B)}) = 0.62$$

$$P(E = yes|Data_5^{(B)}) = 0.83$$

$$P(E = yes|Data_6^{(B)}) = 0.76$$

Since 5 *is the smallest value of* i *such that*

$$P(E = yes|Data_i^{(B)}) > threshold_2 = 0.8,$$

the days to detection is equal to 5 *at a false alarm rate of* 2. *We do not obtain a point on the AMOC curve for a false alarm rate of* 3. *Since* 3 *is the smallest value of* i *such that*

$$P(E = yes|Data_i^{(B)}) > threshold_4 = 0.72,$$

the days to detection is equal to 3 *at a false alarm rate of* 4.

Finally, to obtain the mean days to detection at an annual false alarm rate of r, we compute the average value of the days to detection at an annual false alarm rate of r, where the average is taken over all outbreaks.

PC, PCS, PCT, and PCTS not only detect whether an outbreak is present, but also the type of outbreak. For example, PC not only returns $P(E = yes|Data_i^{(B)})$, but also $P(O = flu|Data_i^{(B)})$. To produce a curve showing how well an influenza outbreak, for example, was detected, the same procedure would be done except that i_r would be the smallest value of i such that

$$P(O = flu|Data_i^{(B)}) > threshold_r. \tag{4.2}$$

4.2.3 Statistical Significance

The performances of some systems were further compared using significance testing (in frequentist terms), which is equivalent to computing the probability that one system's average time to detection is greater than that of another systems (in Bayesian terms). I will now discuss the methodology I used to determine statistical significance.

Suppose we want to compare two systems, $System_1$ and $System_2$, which detect the same set of outbreaks. For a given false alarm rate f, we can analyze the significance of the results using a paired observation t-test. That is, for false alarm rate f, we let

$$\mu_1^{(f)} \text{ be the mean time to detection for } System_1$$
$$\mu_2^{(f)} \text{ be the mean time to detection for } System_2$$

$$\mu^{(f)} = \mu_1^{(f)} - \mu_2^{(f)}.$$

Suppose we are interesting in rejecting the null hypothesis that $System_1$ has a smaller mean time to detection than $System_2$. Then we want see if we can reject that $\mu_1^{(f)} \leq \mu_2^{(f)}$ in favor of $\mu_1^{(f)} > \mu_2^{(f)}$. Our hypotheses are therefore

$$\begin{aligned} H_0^{(f)} &: \mu^{(f)} \leq 0 \\ H_A^{(f)} &: \mu^{(f)} > 0. \end{aligned}$$

The test statistic is then

$$t^{(f)} = \frac{\bar{d}^{(f)}}{s^{(f)}/\sqrt{n}} \tag{4.3}$$

where

$$s^{(f)} = \sqrt{\frac{\sum_{i=1}^{n}(d_i^{(f)} - \bar{d}^{(f)})^2}{n-1}}, \tag{4.4}$$

n is the number of outbreaks, $d_i^{(f)}$ is the difference in detection times for the ith outbreak, $\bar{d}^{(f)}$ is the average difference in detection times, and the test statistic has the t distribution with $n-1$ degrees of freedom. Using the test statistic in Equality 4.3, we compute the p-value $p^{(f)}$ of the result.

Suppose instead we do a Bayesian analysis. In the Bayesian framework, we first assume that the difference X in the detection times is normally distributed with unknown mean and unknown precision (the precision is one divided by the variance). We represent our belief concerning the unknown mean and unknown precision with the random variables A and R respectively. We assume that our prior belief concerning the value of R is represented by a Gamma probability density function and that our prior belief concerning the value of A is represented by a conditional Normal density function. Conditional on the $Data$, we then compute the posterior probability distributions of R, A, and X. (Neapolitan, 2004, p. 405) discusses that we can model prior ignorance concerning the mean and precision by assuming that the prior density function of R is the improper density function $1/r$ and the prior density function of A is the improper uniform density function over the whole real line. Given that we do this, I obtain results in Appendix A which imply that

$$P(H_A^{(f)}|Data) = P(\mu_1^{(f)} > \mu_2^{(f)}|Data) = 1 - p^{(f)}. \tag{4.5}$$

where $p^{(f)}$ is the p-value obtained using Equation 4.3. The test was introduced using frequentist terminology because that terminology seems to be more well-known. However, Equality 4.5 (Bayesian terminology) will be used when showing results because that terminology seems more intuitive.

I deviated from actually performing a paired-observation t-test in two ways. First, due to the large sample size, I assumed that the sample standard deviation was about equal to the actual standard deviation. So, instead of using a t-test and Equality 4.4, I was able to use the Z-test with known variance. That is, the test statistic was

$$z^{(f)} = \frac{\bar{d}^{(f)}}{\sigma^{(f)}/\sqrt{n}}$$

where

$$\sigma^{(f)} = \sqrt{\frac{\sum_{i=1}^{n}(d_i^{(f)} - \bar{d}^{(f)})^2}{n}}.$$

Furthermore, because the systems were run separately, I was only able to obtain the individual standard deviations, $\sigma_1^{(f)}$ and $\sigma_2^{(f)}$, for each system. However, since

$$\left(\sigma^{(f)}\right)^2 = \left(\sigma_1^{(f)}\right)^2 + \left(\sigma_2^{(f)}\right)^2 - 2Cov(D_1, D_2)$$

where D_i is a random variable representing the detection time for $System_i$, the variance can be approximated as follows:

$$\left(\sigma^{(f)}\right)^2 \approx \left(\sigma_1^{(f)}\right)^2 + \left(\sigma_2^{(f)}\right)^2.$$

If the random variables were independent, this approximation would be exact. However, it is more reasonable to assume that the random variables are positively correlated than that they are independent. That is, if a given outbreak were harder to detect for one system then in would be harder to detect for the other system. Given this assumption of positive correlation, the approximation is expected to be an upper bound on the actual variances, which implies that the p-value obtained is an upper bound on the actual p-value. This means that the probability that $System_1$ has a larger mean time to detection than $System_2$ is *at least as large* as the value obtained (see Equality 4.5).

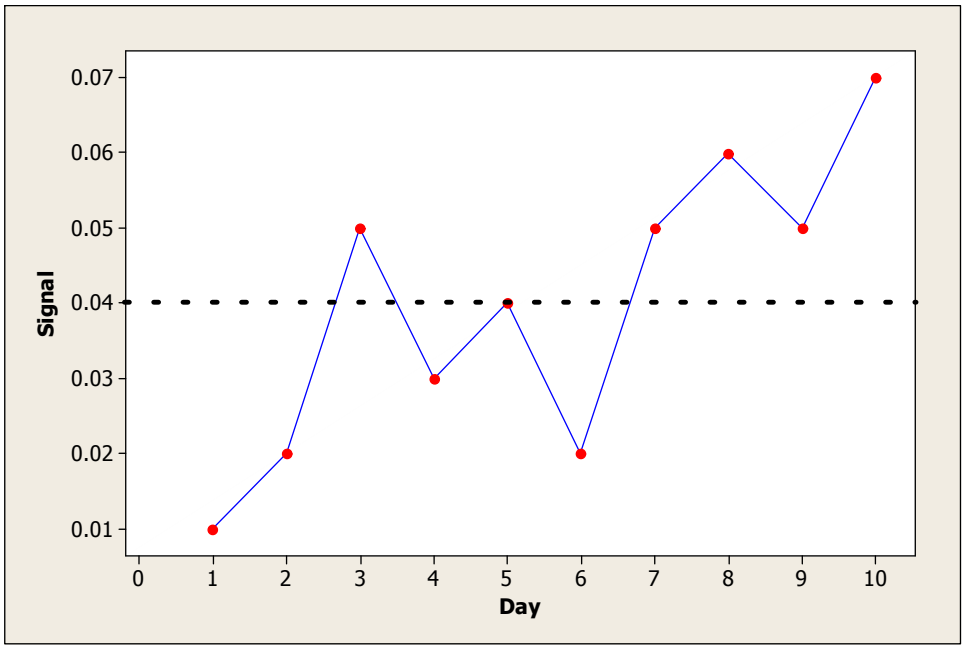


Figure 4.3: On the 7th day the signal exceeds 0.4 and stays at or above that level.

4.2.4 AMOC-M Curves

We are not only interested in how early a system can detect an outbreak, but also in how early the system maintains the detection of an outbreak. AMOC-M curves were used to evaluate the latter. An **AMOC-M curve (AMOC-Maintain curve)**[1] is just like an AMOC curve except that the y-axis plots the average of the time at which an outbreak signal is detected and maintained thereafter. For example, if the threshold is .04, and the sequence of signals is [.01, .02, .05, .03, .04, .02, **.05**, .06, .05, .07], then the time at which the signal is maintained is 7 because on the 7th day the probability is .05, which exceeds .04, and it stays at or above .04 after that. This example is shown in Figure 4.3.

Note that an AMOC-M curve represents a family of performance measures. That is, in general there would be a parameter t, whose value is the time period over which the signal must stay above the threshold. For example, if $t = 3$, the signal would need to be

[1] To my knowledge, an AMOC-M curve has not been previously defined. It is innovative in this thesis.

maintained for 3 days, while if $t = 5$ the signal would need to be maintained for 5 days. In my experiments, I required that the signal be maintained for the duration of the simulation, which was 15 days. For example, if the signal was below the threshold on the 6th of the simulation and stayed above the threshold from the 7th day until the 15th day, then I would say that the signal was detected and maintained on the 7th day. When we set t to the outbreak duration, the AMOC-M curve measures how well a method can help public health officials in their action/investigation phase after the initial detection is made. For example, if the signal is maintained long enough, then it may suggest that the outbreak is stable enough to warrant sending investigators into the community to better understand the outbreak. If the signal does not persist very long, then it may suggest that the outbreak is too tenuous to warrant an extensive investigation. However, my requirement that the signal must be maintained for the duration of the outbreak is somewhat severe. In practice, public health officials must typically make the decision whether to respond to the outbreak signal long before the outbreak has reached its peak, and therefore they might need to make their decision by the time the signal has been maintained for 5 days or even 3 days. So assuming t is equal to the outbreak duration is not suitable for measuring the usefulness of a method for detection. To solve this problem, we should consider using a fixed value of t instead. However, when using a fixed value of t with an AMOC-M curve, it is important to further investigate how we determine a practical value of t relative to various potential applications of AMOC-M curve such as the early and reliable detection of an outbreak, the decision making of public health officials, and the evaluation of a detection method by researchers.

4.2.5 Subregion Detection

The systems that perform spatial analysis not only detect an outbreak, but also determine the spatial subregion in which the outbreak is occurring. There are three ways in which the accuracy of the subregion region detected was analyzed, namely the overlap coefficient, the precision, and the spatial recall. Let S be the correct subregion (i.e. the subregion in which the outbreak was injected), T be the detected subregion, and # returns the number of zip

codes in a subregion. The values of these measures are then as follows:

$$overlap\ coefficient(S,T) = \frac{\#(S \cap T)}{\#(S \cup T)}$$

$$precision(S,T) = \frac{\#(S \cap T)}{\#(T)}$$

$$spatial\ recall(S,T) = \frac{\#(S \cap T)}{\#(S)}.$$

The overlap coefficient is 0 if and only if two the two subregions do not intersect, while it is 1 if and only if they are the same subregion. The value of the precision is 0 if and only if the two subregions do not intersect, while it is 1 if and only if $T \subseteq S$. The value of spatial recall is 0 if and only if the two subregions do not intersect, while it is 1 if and only if $S \subseteq T$.

4.3 EXPERIMENTS

Next I show results of experiments testing the four hypotheses of this thesis.

4.3.1 Method

As discussed in Section 4.2.1, Allegheny County, Pennsylvania was modeled using a 16×16 grid, and each grid element is one cell. Both influenza and *Cryptosporidium* outbreaks were simulated in rectangular subregions of that county. The properties of the simulations were as follows.

1. **Epidemic Curve Function:** I simulated outbreaks that increase according to linear, quadratic, and cubic functions.
2. **Outbreak Severity**: For each cell, I determined the mean and standard deviation σ_{cell} of the number of real ED visits during a one-year background period when it was assumed no outbreak was occurring. The period chosen was the entire calendar year 2004. An outbreak was simulated by injecting simulated ED visits into the background period of real data. The data for each simulation consisted of the new injected ED visits plus the ED data in the background period. So the data was semi-synthetic. The severity level of

an outbreak was based on a multiple of σ_{cell}. If, for example, the severity level was based on $2\sigma_{cell}$, the average daily number of injected ED visits was $2\sigma_{cell}$. The duration of all outbreaks was set equal to 30 days. Therefore, if the severity level was based on $2\sigma_{cell}$, the total number of ED visits injected into the cell during the simulation was given by

$$tot_{cell} = 30 \times 2\sigma_{cell}.$$

The multiples of σ_{cell} used for each epidemic curve function were as follows:

Function	Severity Level 1	Severity Level 2
linear	$1.5\sigma_{cell}$	$2\sigma_{cell}$
quadratic	$2\sigma_{cell}$	$2.5\sigma_{cell}$
cubic	$2.5\sigma_{cell}$	$3\sigma_{cell}$

I used larger multiples in the case of quadratic and cubic functions because otherwise it would have taken too long for the number of injections to reach a detectable level.

3. **Daily Increase:** I assumed that half of the injected ED visits occurred during the first half of the outbreak. In the case of outbreaks that methodically exhibited a linear increase in outbreak cases, we would assume that $\triangle$ of them occur on day one of the outbreak, $2\triangle$ occur on day two, and so on. The value of $\triangle$ can therefore be determined by solving

$$\triangle + 2\triangle + \cdots + \frac{30}{2}\triangle = \frac{tot_{cell}}{2},$$

which is the same as

$$\frac{(\frac{30}{2})(\frac{30}{2}+1)}{2}\triangle = \frac{tot_{cell}}{2}.$$

To simulate an outbreak that methodically exhibited a linear increase in outbreak cases, $\triangle$ new outbreak cases would be injected into the background data on day one of the simulation period, $2\triangle$ on day two, and so on. Figure 4.4 shows the total ED visits for such a simulated outbreak.

The formula for determining Δ for an outbreak that shows a quadratic increase is

$$1^2\triangle + 2^2\triangle + \cdots + \left(\frac{30}{2}\right)^2\triangle = \frac{tot_{cell}}{2},$$

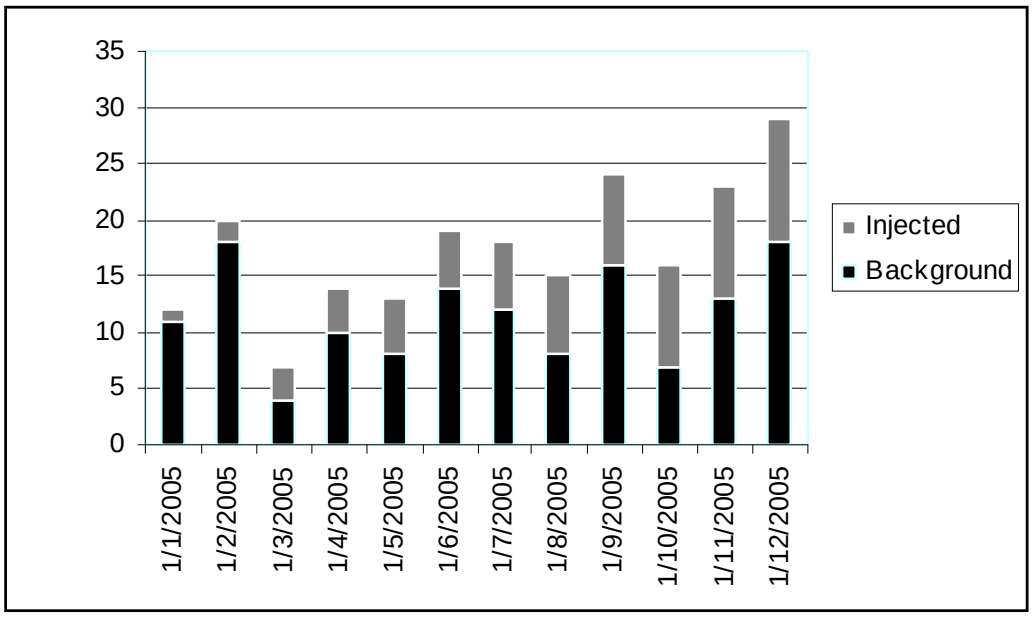


Figure 4.4: A simulated outbreak.

which is the same as

$$\frac{\left(\frac{30}{2}\right)\left(\frac{30}{2}+1\right)\left(2\left(\frac{30}{2}\right)+1\right)}{6}\Delta = \frac{tot_{cell}}{2}.$$

Finally, the formula for determining Δ for an outbreak that shows a cubic increase is

$$1^3\triangle + 2^3\triangle + \cdots + \left(\frac{30}{2}\right)^3 \triangle = \frac{tot_{cell}}{2},$$

which is the same as

$$\left(\frac{\left(\frac{30}{2}\right)\left(\frac{30}{2}+1\right)}{2}\right)^2 \Delta = \frac{tot_{cell}}{2}.$$

To simulate an outbreak that methodically exhibited a quadratic increase, $\triangle$ new outbreak cases would be injected into the background data on day one of the simulation period, $2^2\triangle$ on day two, and so on. To simulate an outbreak that methodically exhibited a cubic increase, $\triangle$ new outbreak cases would be injected into the background data on day one of the simulation period, $2^3\triangle$ on day two, and so on.

To force daily fluctuations, I deviated from simply making the number of new cases on day t equal to $t\Delta$ (linear case), $t^2\Delta$ (quadratic case), or $t^3\Delta$ (cubic case). Rather, I made the injection curve **multimodal** as follows. In half the simulations, on even numbered days I made the number of new cases 25% of the previous day's number, and in the other half I made it 50% of the previous day's number. I imposed daily fluctuations so I could evaluate the detection maintenance capability of the systems. If PC first detected an outbreak on day t when the number of injections was, for example, 100, it is likely that it would not detect it on day $t+1$ if the number of injections was only 50. However, since PCT would be looking at the data from both day t and day $t+1$ it seems likely that it would maintain the detection signal on day $t+1$. Perhaps it would be more realistic to generate the number of daily injections using Poisson distributions, where each day the mean of the distribution increases. However, my purpose was to make certain that I used simulations that challenged the detection maintenance capability of the systems, and forcing daily fluctuations guaranteed that my simulations exhibited day fluctuations. If I used Poisson distributions instead, I may have had to do many more simulations to get a good sample of the kinds of outbreaks I needed to investigate.

4. **Chief Complaint:** To determine the chief complaint of each injected case, the chief complaint was generated at random using a probability distribution Q of the chief complaints given the disease (influenza or *Cryptosporidium*) whose outbreak was simulated. Recall that PC contains a probability distribution P of the chief complaints given each of the outbreak diseases. In order to test the robustness of the systems, Q was allowed to vary significantly from P.

   To obtain a conditional probability distribution Q, I let the conditional probabilities of the chief complaints in PC be the means of Dirichlet distributions. For example, if an influenza outbreak was being simulated, if p_1, p_2, ..., and p_{54} are the conditional probabilities in PC for each of the chief complaints given influenza, and N is our subjective prior sample size, I set

   $$a_i = p_i N$$

   to obtain parameters for a Dirichlet distributions. Once such a Dirichlet distribution was developed, I randomly generated a probability distribution Q according to this distrib-

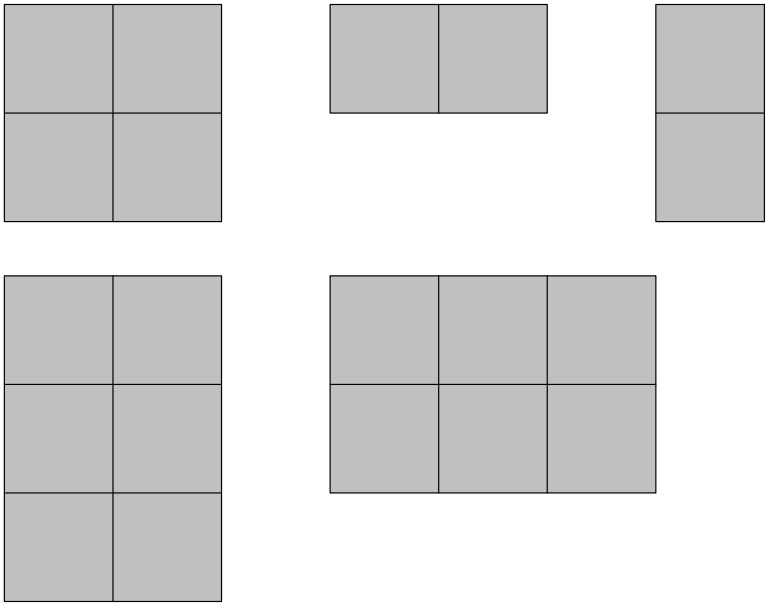


Figure 4.5: The shapes of the injected subregions.

ution. As N increases, the likelihood of Q being similar to the conditional probability distribution P in PC increases. In these experiments I used $N = 5$. Ten different probability distributions were generated.

5. **Outbreak Subregions**: Outbreaks that occur in rectangles that are 2 cells by 1 cell, 2 cells by 2 cells, and 3 cells by 2 cells were developed. The 2 by 1 rectangles and 3 by 2 rectangles could go either north-south or east-west. Figure 4.5 shows the shapes of the various rectangles. Not all cells had sufficient background ED visits to qualify for containing an outbreak. All those cells that had $\sigma_{cell} < 1$ were eliminated. Each 2 by 1 rectangle, for example, was generated by randomly choosing two contiguous cells. If either cell had $\sigma_{cell} < 1$, a new rectangle would be generated. This was done until a 2 by 1 rectangle was obtained that did not have $\sigma_{cell} < 1$ in either cell.

For each outbreak disease, for each type of increase (linear, quadratic, cubic), I performed 4 simulations with each of the 10 probability distributions. Therefore, there were a total of 40 simulations for each type of increase, and a total of 120 simulations total for each

outbreak disease. The properties of the 40 outbreaks for each outbreak disease and type of increase were determined as follows:

Variable	Values	# Occurrences of Each Value	Total # Occurrences
Prob. Dist.	$1, 2, 3, 4, 5, 6, 7, 8, 9, 10$	4	40
Month	$1 - 12$	$3 - 4$	40
Day	$1 - 30$	$1 - 2$	40
Severity	$1, 2$	20	40
Percent	$25\%, 50\%$	20	40
Subregion	4 each of types 2 by 1, 2 by 2, and 3 by 2	$3 - 4$	40

The *Month* and *Day* variables determined the starting date for the simulated outbreak. For example, if $Month = 3$ and $Day = 5$, the simulated outbreak started on March 5 of the one-year background period.

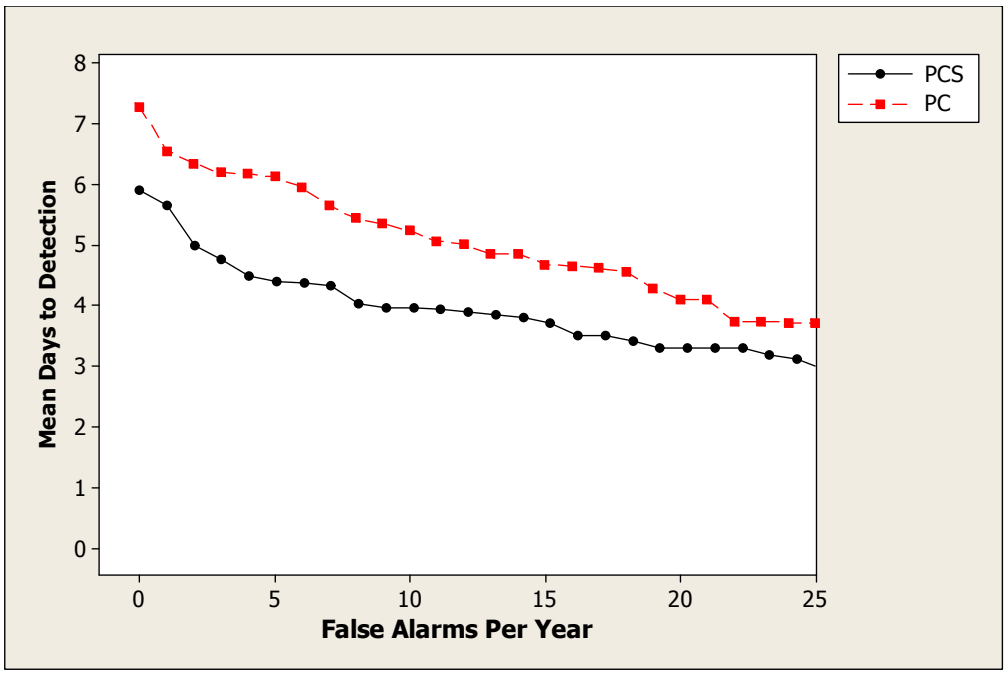
For each variable, a list of the 40 occurrences was created. To develop each outbreak, a value of each variable in the table above was sampled at random without replacement from each list.

Example 4.2. *Suppose the sampled values were as follows: Prob. Dist.* $= 4$*; Month* $= 7$*; Day* $= 23$*; Severity* $= 2$*; Percent* $= 25\%$*; Subregion* $= 2$ *by* 1*. Then the 4th probability distribution was used, injections started on date* $7/23/2004$ *of the background time period, the severity level was* 2*, the percent factor was* 25%*, and the injections occurred in a* 2 *by* 1 *cell subregion.*

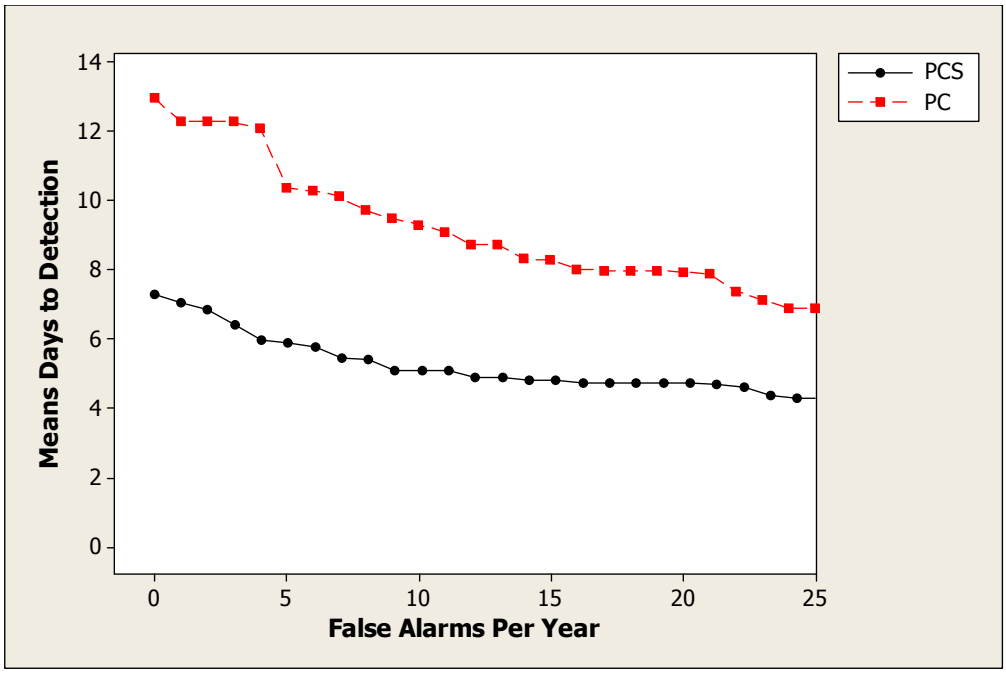
The AMOC curves in all the results that follow show the abilities of PC, PCS, PCT, and PCTS to detect the specific disease in the outbreaks. That is, Equation 4.2 was used to produce the AMOC curves.

4.3.2 Results of Testing Hypothesis 1 (PCS Improves PC)

4.3.2.1 AMOC Curves Figure 4.6 shows AMOC curves comparing the detection performance of PCS and PC. In the cases of both *Cryptosporidium* and influenza outbreaks, they indicate that PCS performs better.



(a) *Cryptosporidium* outbreaks



(b) Flu outbreaks

Figure 4.6: AMOC curves comparing the detection performance of PCS and PC.

FAR	$P(\mu_{PCc} > \mu_{PCSc})$	$P(\mu_{PCf} > \mu_{PCSf})$
0	1	1
5	1	1
10	1	1
15	1	1

Table 4.1: At various false alarm rates (FAR), the posterior probabilitiy that PCS has a smaller mean day to detection than PC.

4.3.2.2 Significance Testing of Detection Power Table 4.1 shows the posterior probability that PCS has a smaller mean day to detection than PC at various false alarm rates (FAR). In this and similar tables the value 1 denotes that the probability is greater than 0.99999. Furthermore, in this table the following notation was used to refer to the systems:

PCc: PC detecting *Cryptosporidium* outbreaks.

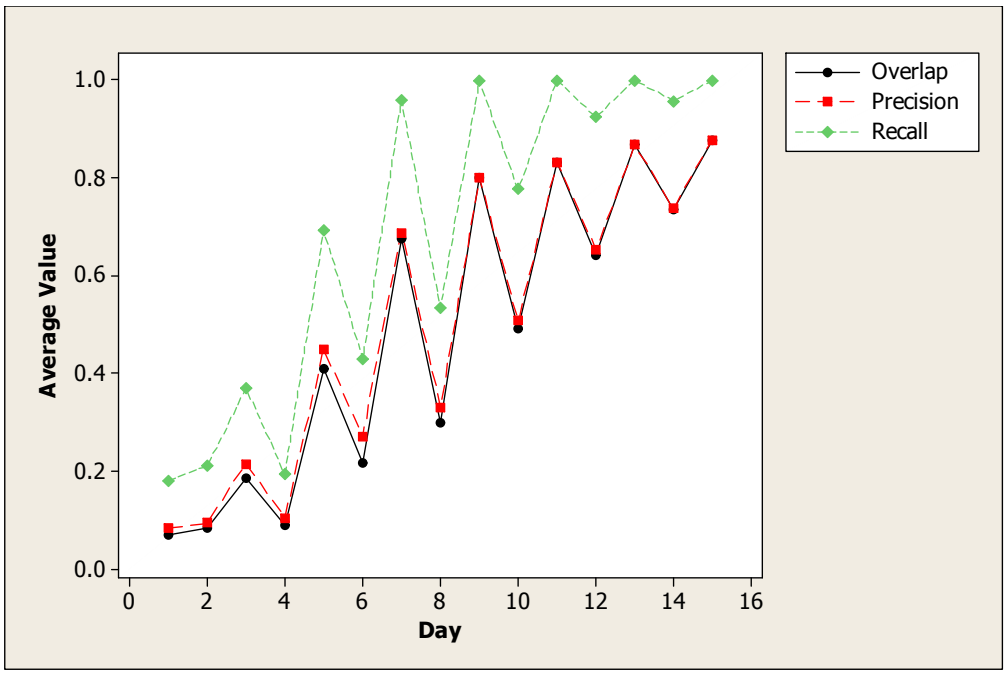
$PCSc$: PCS detecting *Cryptosporidium* outbreaks.

PCf: PC detecting influenza outbreaks.

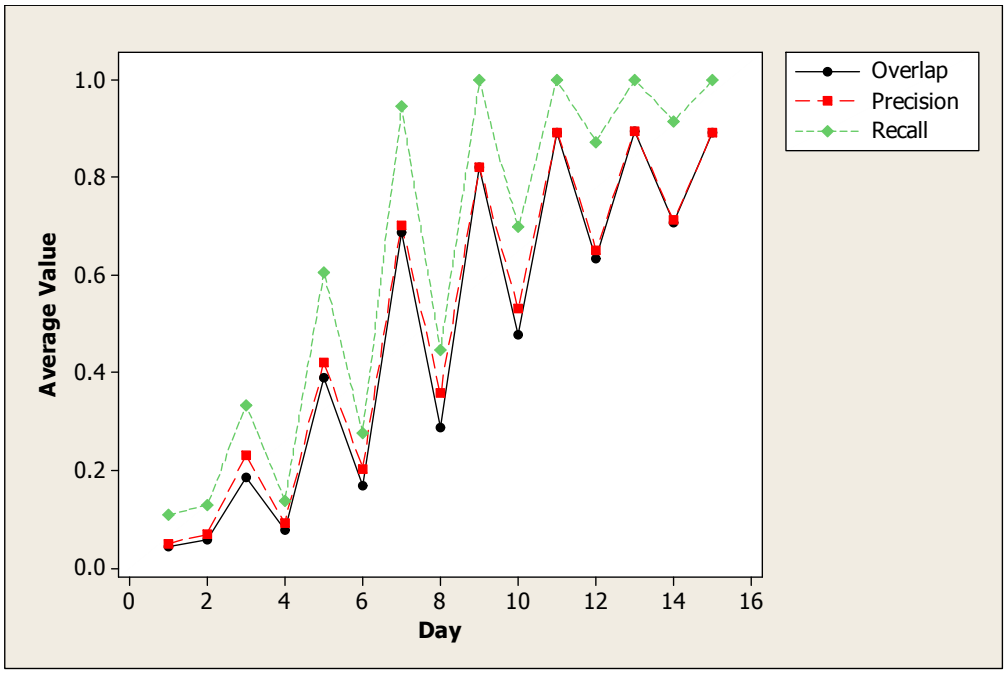
$PCSf$: PCS detecting influenza outbreaks.

These results support the first part of hypothesis 1, namely that PCS is an improvement on PC in that it will have a smaller mean time to detection at most false alarm rates.

4.3.2.3 Subregion Detection Figure 4.7 shows the average values of the overlap coefficient, precision, and spatial recall for PCS. The subregion S that maximized $P(Data|SUB = S)$ was considered to be the subregion detected by PCS. Note that this is the same subregion with maximum posterior probability if we assume that all subregions have the same prior probability. As expected, the values fluctuate up and down because we injected fewer cases on alternate days. However, by the 7th day, all values are fairly good (in particular spatial recall) on the high days. These results support the second part of hypothesis 1, namely



(a) *Cryptosporidium* outbreaks



(b) Flu outbreaks

Figure 4.7: The average values of the overlap coefficient, precision, and spatial recall for PCS.

FAR	$P(\mu_{PCc} > \mu_{PCTc})$	$P(\mu_{PCf} > \mu_{PCTf})$
0	0.9038	0.1291
5	0.6791	0.9132
10	0.0877	0.9087
15	0.0198	0.5630

Table 4.2: At various false alarm rates (FAR), the posterior probabilitiy that PCT has a smaller mean day to detection than PC when.

that PCS can accurately locate the subregion in which an outbreak is occurring when the outbreak is restricted to a subregion of the monitored region.

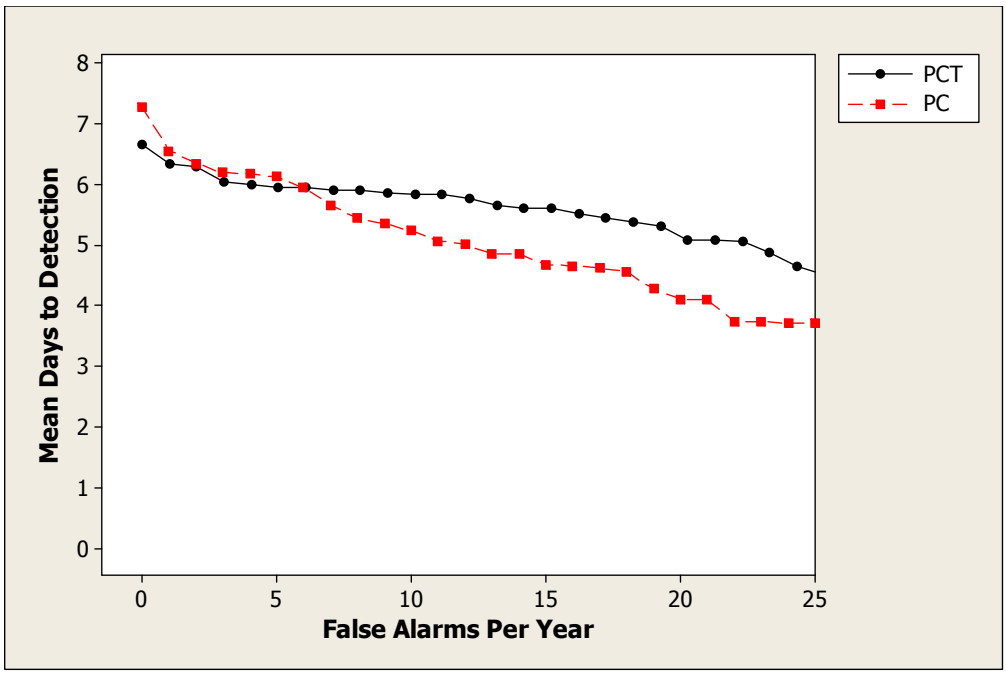
4.3.3 Results of Testing Hypothesis 2 (PCT Improves PC)

4.3.3.1 AMOC Curves Figure 4.8 shows AMOC curves comparing the detection performance of PCT and PC. In the case of *Cryptosporidium* outbreaks PCT performed better for small false alarm rates, but worse for large false alarm rates. In the case of influenza outbreaks the performance of the two systems was about the same.

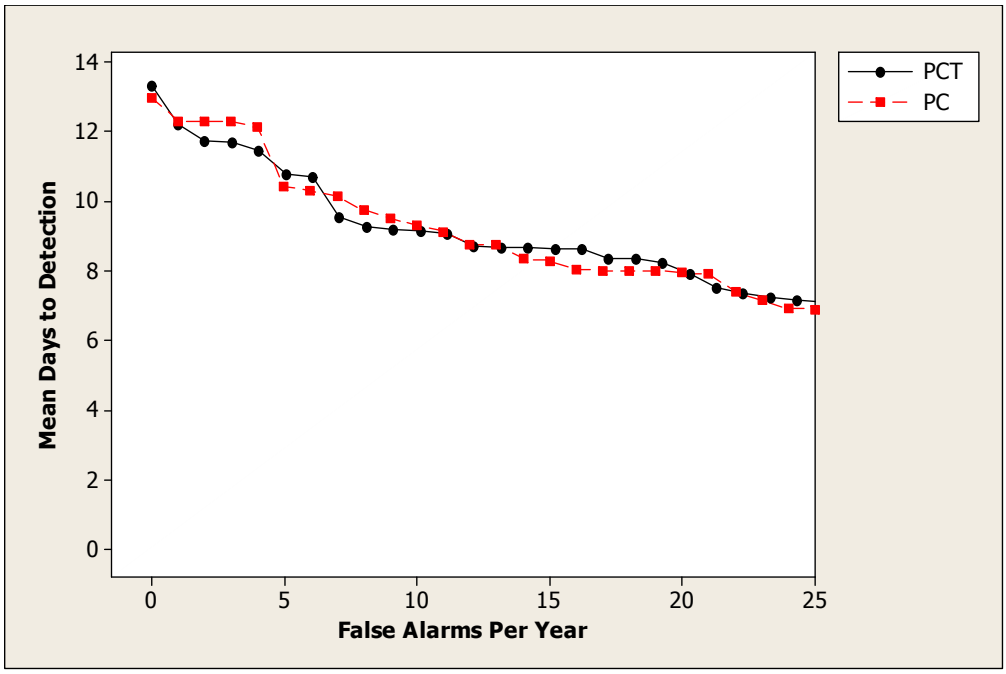
4.3.3.2 Significance Testing of Detection Power Table 4.2 shows the posterior probability that PCT has a smaller mean day to detection than PC at various false alarm rates (FAR). In that table the following notation was used:

PCc: PC detecting *Cryptosporidium* outbreaks.
$PCTc$: PCT detecting *Cryptosporidium* outbreaks.
PCf: PC detecting influenza outbreaks.
$PCTf$: PCT detecting influenza outbreaks.

In five out of 8 cases, the results in Table 4.2 favor PCT. Furthermore, according to traditional standards for statistical significance, only one result is significant. Namely for $FAR = 15$, $P(\mu_{PCc} > \mu_{PCTc}) = 0.0198$. Ordinarily, we would not run a detection system



(a) *Cryptosporidium* outbreaks



(b) Flu outbreaks

Figure 4.8: AMOC curves comparing the detection performance of PCT and PC.

FAR	$P(\nu_{PCc} > \nu_{PCTc})$	$P(\nu_{PCf} > \nu_{PCTf})$
0	1	0.9990
5	1	1
10	1	1
15	1	1

Table 4.3: At various false alarm rates (FAR), the posterior probabilitiy that PCT has a smaller mean day to maintaining detection than PC.

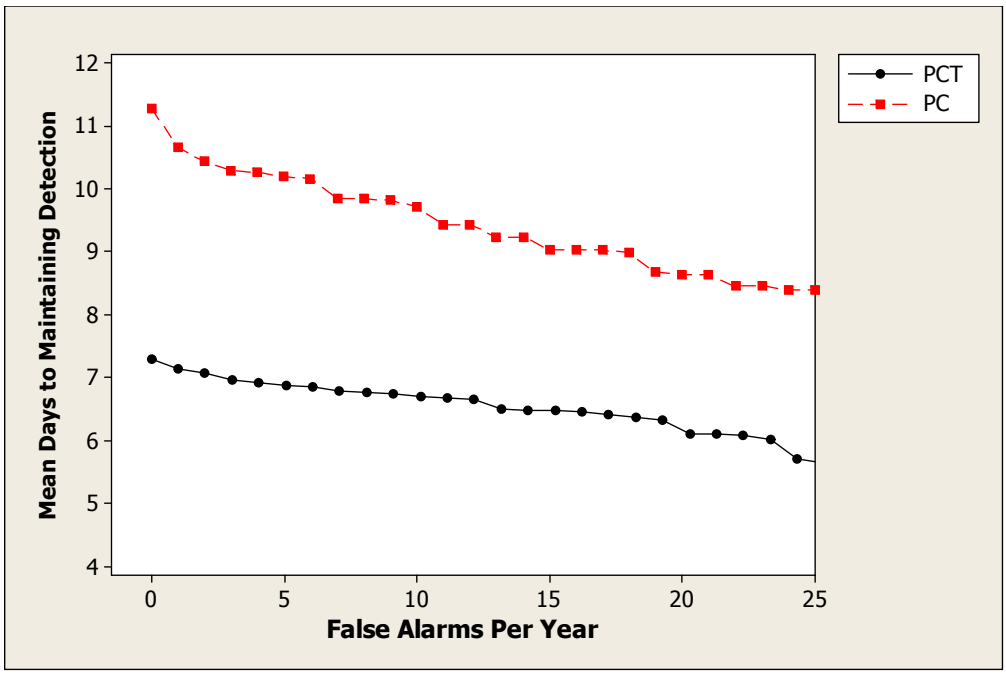
with an annual false alarm rate equal to 15. So this result is not very significant either. We conclude that the results support the second part of hypothesis 2, which is that initial detection performance is not compromised when we use PCT.

4.3.3.3 AMOC-M Curves Figure 4.9 shows AMOC-M curves comparing the detection maintenance performance of PCT and PC. For both *Cryptosporidium* and influenza outbreaks, the performance of PCT is superior to that of PC for all false alarm rates.

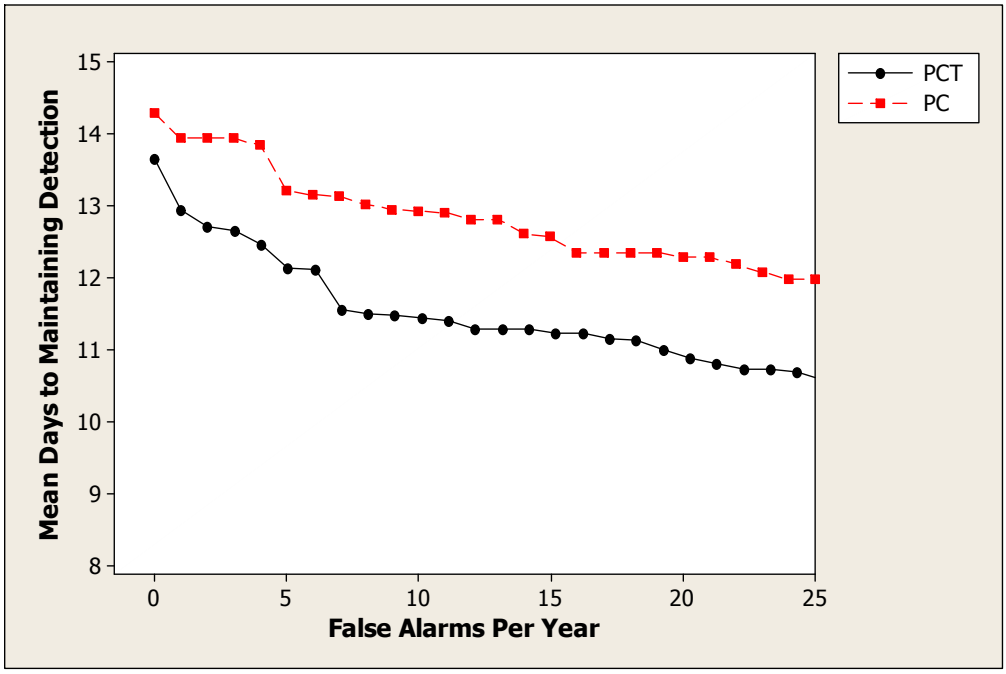
4.3.3.4 Significance Testing of Detection Maintenance Power Table 4.3 shows the posterior probability that PCT has a smaller mean day to maintaining detection than PC at various false alarm rates (FAR). In that table the following notation was used to refer to the systems:

PCc: PC detecting *Cryptosporidium* outbreaks.
$PCTc$: PCT detecting *Cryptosporidium* outbreaks.
PCf: PC detecting influenza outbreaks.
$PCTf$: PCT detecting influenza outbreaks.

These results support the first part of hypothesis 2, which is that PCT is more stable than PC in that once an outbreak is detected, PCT is better at maintaining the detection signal on future days.



(a) *Cryptosporidium* outbreaks



(b) Flu outbreaks

Figure 4.9: AMOC-M curves comparing the detection maintenance performance of PCT and PC.

FAR	$P(\mu_{PCTc} > \mu_{PCTSc})$	$P(\mu_{PCTf} > \mu_{PCTSf})$
0	0.9999	1
5	1	1
10	1	1
15	1	1

Table 4.4: At various false alarm rates (FAR), the posterior probabilitiy that PCTS has a smaller mean day to detection than PCT.

4.3.4 Results of Testing Hypothesis 3 (PCTS Improves PCT)

4.3.4.1 AMOC Curves Figure 4.10 shows AMOC curves comparing the detection performance of PCTS and PCT. These curves illustrate the superior detection performance of PCTS.

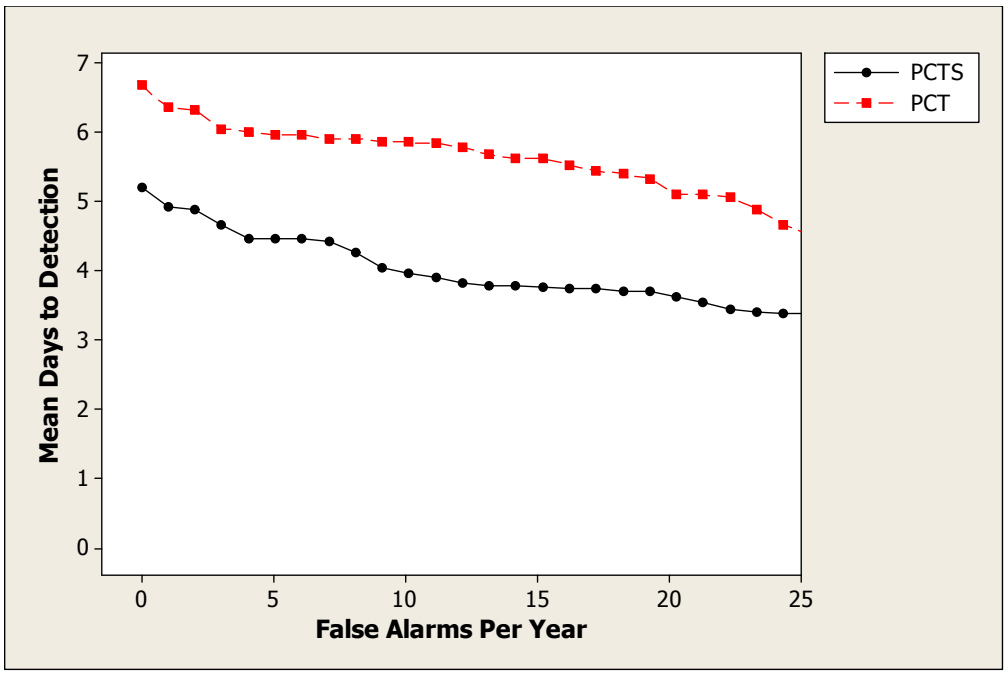
4.3.4.2 Significance Testing of Detection Power Table 4.4 shows the posterior probability that PCTS has a smaller mean day to detection than PCT at various false alarm rates (FAR). In that table the following notation was used to refer to the systems:

$PCTc$: PCT detecting *Cryptosporidium* outbreaks.
$PCTSc$: PCTS detecting *Cryptosporidium* outbreaks.
$PCTf$: PCT detecting influenza outbreaks.
$PCTSf$: PCTS detecting influenza outbreaks.

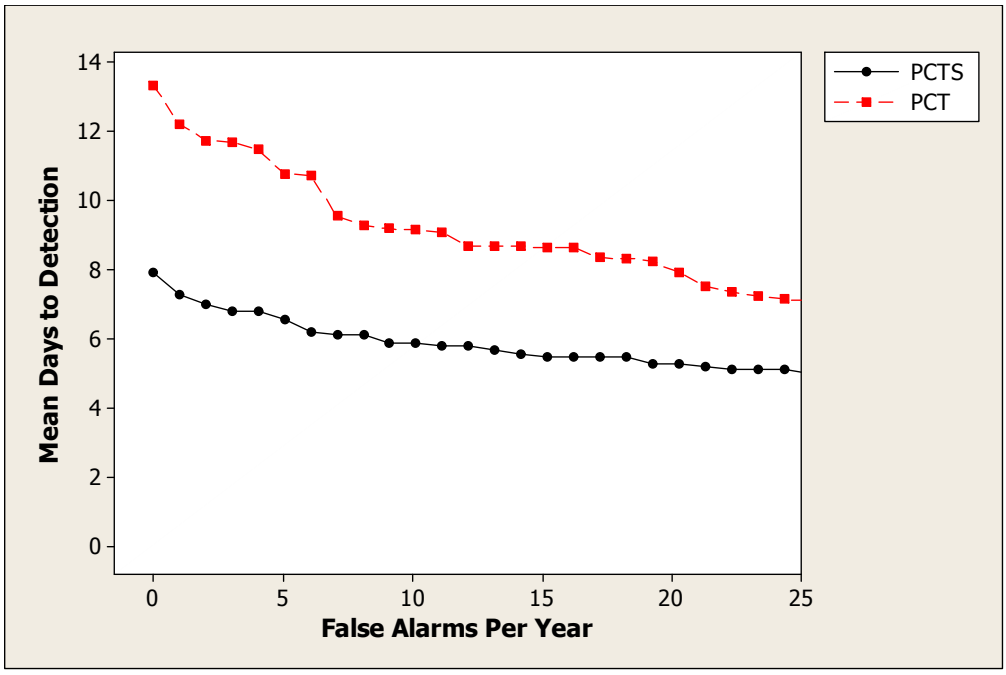
These results support hypothesis 3, which states that PCTS is an improvement over PCT in that it will have a smaller mean time to detection at most false alarm rates.

4.3.5 Results of Testing Hypothesis 4 (PCTS Improves PCS)

4.3.5.1 AMOC Curves Figure 4.11 shows AMOC curves comparing the detection performance of PCTS and PCS. In the case of the *Cryptosporidium* outbreaks, PCTS performed

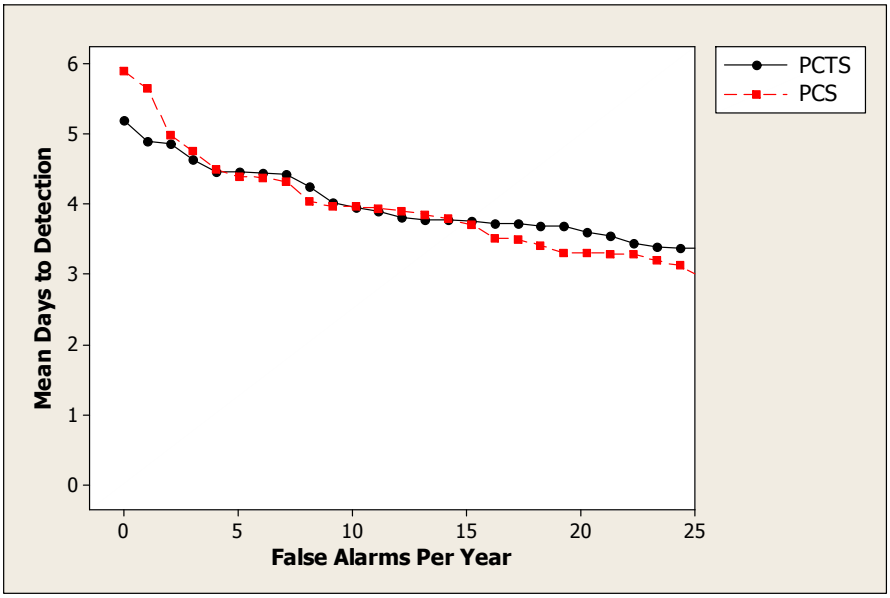


(a) *Cryptosporidium* outbreaks

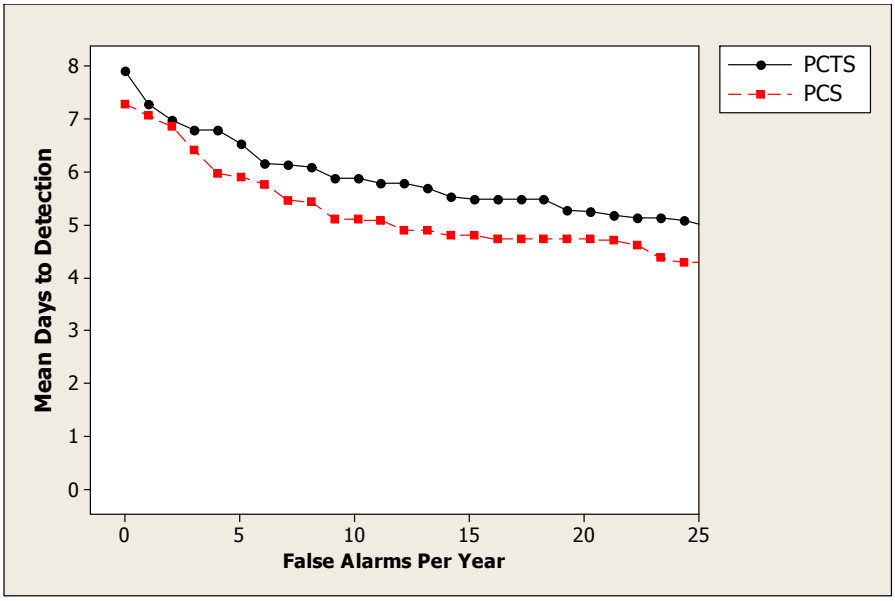


(b) Flu outbreaks

Figure 4.10: AMOC curves comparing the detection performance of PCTS and PCT.



(a) *Cryptosporidium* outbreaks



(b) Flu outbreaks

Figure 4.11: AMOC curves comparing the detection performance of PCTS and PCS.

FAR	$P(\mu_{PCTc} > \mu_{PCTSc})$	$P(\mu_{PCTf} > \mu_{PCTSf})$
0	0.9647	0.0402
5	0.3909	0.0406
10	0.5218	0.1192
15	0.4332	0.0662

Table 4.5: At various false alarm rates (FAR), the posterior probabilitiy that PCTS has a smaller mean day to detection than PCT.

slightly better than PCS early in the outbreak, and performed slightly worse very late in the outbreak. In the case of the influenza outbreaks, PCTS performed slightly worse than PCS.

4.3.5.2 Significance Testing of Detection Power Table 4.5 shows the posterior probability that PCTS has a smaller mean day to detection than PCT at various false alarm rates (FAR). In that table the following notation was used to refer to the systems:

$PCTc$: PCT detecting *Cryptosporidium* outbreaks.

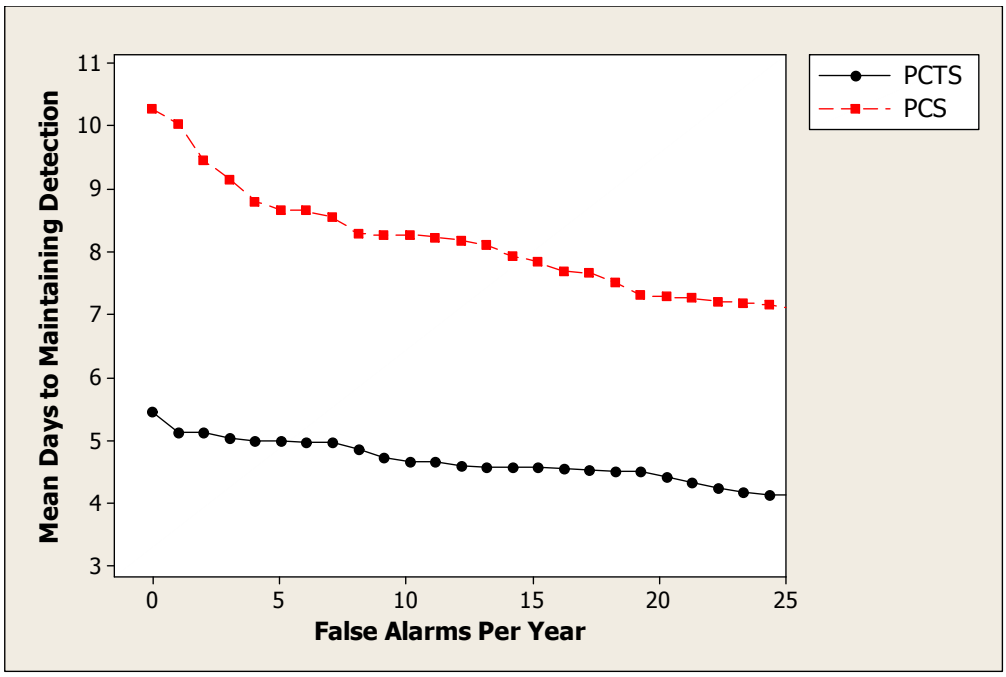
$PCTSc$: PCTS detecting *Cryptosporidium* outbreaks.

$PCTf$: PCT detecting influenza outbreaks.

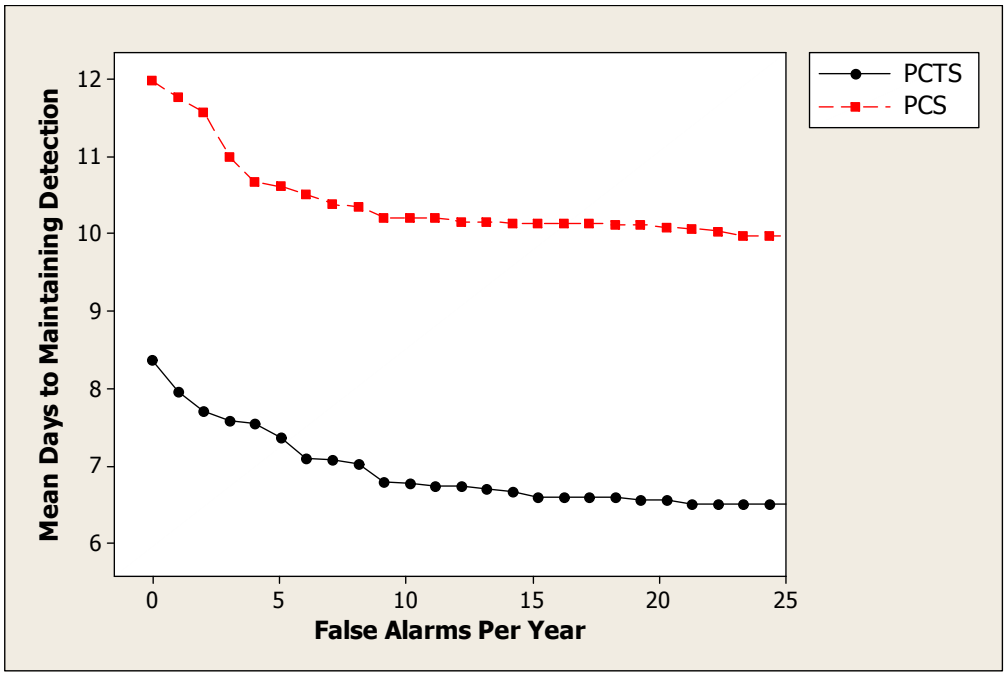
$PCTSf$: PCTS detecting influenza outbreaks.

In the case of *Cryptosporidium*, only one result ($FAR = 0$) would be considered significant by traditional standards, and that result favors PCTS. In the case of influenza, two results ($FAR = 0$ and $FAR = 5$) would be considered significant by traditional standard, and those result favors PCS. We conclude that PCTS may show slightly degraded performance relative to PCS, but such degradation is not strongly supported by the results.

4.3.5.3 AMOC-M Curves Figure 4.12 shows AMOC-M curves comparing the detection maintenance performance of PCTS and PCS. PCTS's performance is far superior.



(a) *Cryptosporidium* outbreaks



(b) Flu outbreaks

Figure 4.12: AMOC-M curves comparing the detection maintenance performance of PCTS and PCS.

FAR	$P(\nu_{PCSc} > \nu_{PCTSc})$	$P(\nu_{PCSf} > \nu_{PCTSf})$
0	1	1
5	1	1
10	1	1
15	1	1

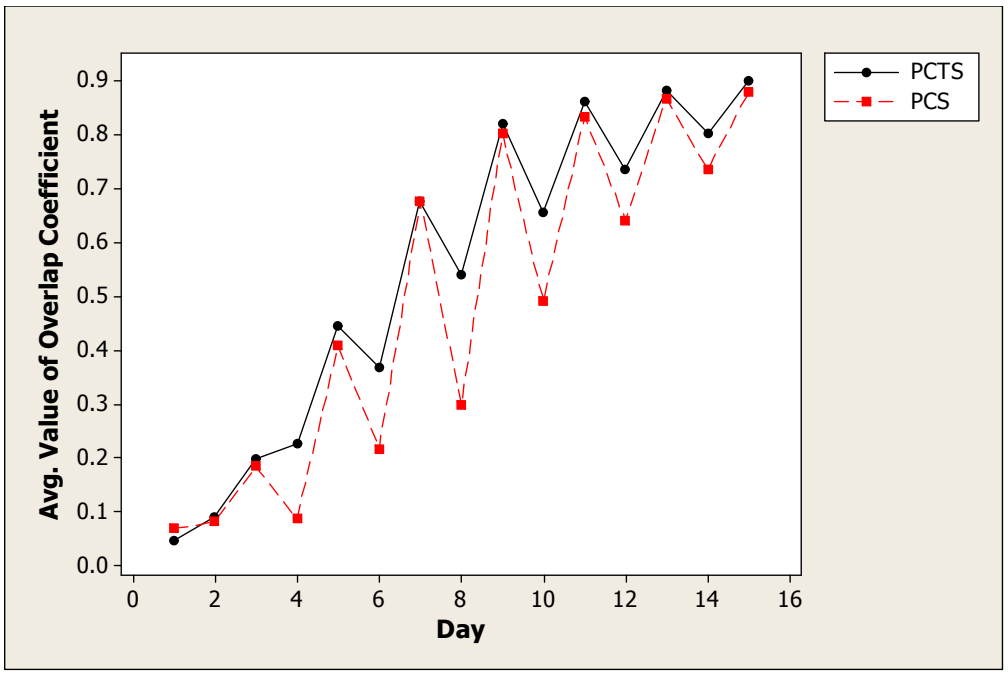
Table 4.6: At various false alarm rates (FAR), the posterior probabilitiy that PCTS has a smaller mean day to maintaining detection than PCS.

4.3.5.4 Significance Testing of Detection Maintenance Power Table 4.6 shows the posterior probability that PCTS has a smaller mean day to maintaining detection than PCS at various false alarm rates (FAR). In that table the following notation was used:

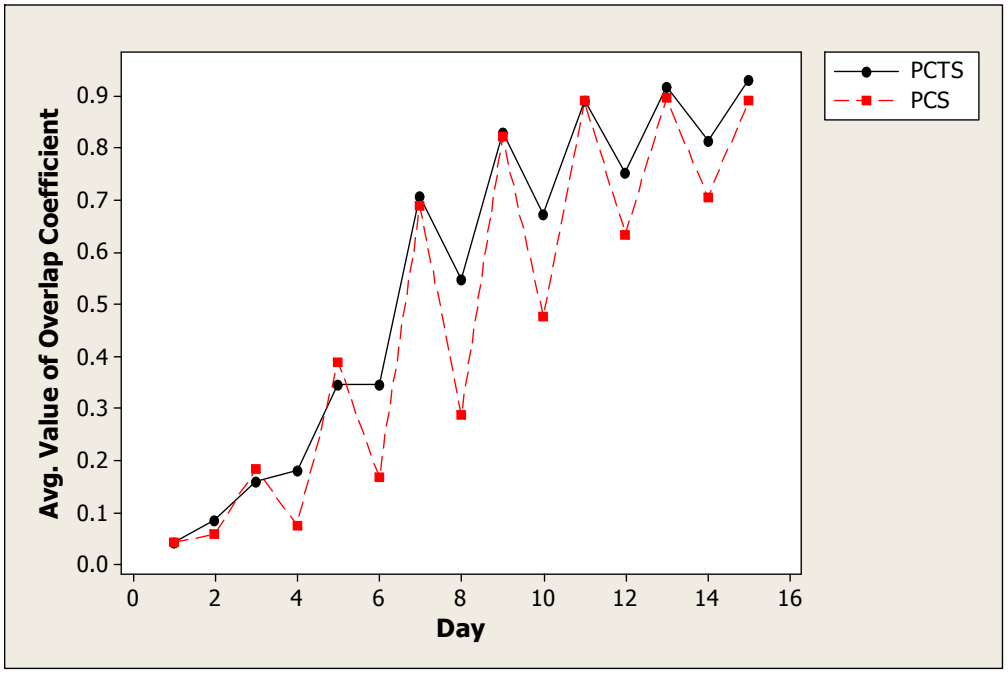
$PCSc$: PCS detecting *Cryptosporidium* outbreaks.
$PCTSc$: PCTS detecting *Cryptosporidium* outbreaks.
$PCSf$: PCS detecting influenza outbreaks.
$PCTSf$: PCTS detecting influenza outbreaks.

These results support the first part of hypothesis 4, which is that once an outbreak is detected, PCTS can better maintain the detection signal on future days.

4.3.5.5 Subregion Detection Figures 4.13, 4.14, and 4.15 show the average values of the overlap coefficient, precision, and spatial recall respectively for PCTS and PCS. In all cases, PCTS and PCS perform about the same on odd numbered days (when a large number of cases was injected), but the performance of PCS degrades much more on even numbered days (when a smaller number of cases was injected) than does the performance of PCTS. So temporal modeling is also beneficial as far as maintaining detection of the correct subregion.

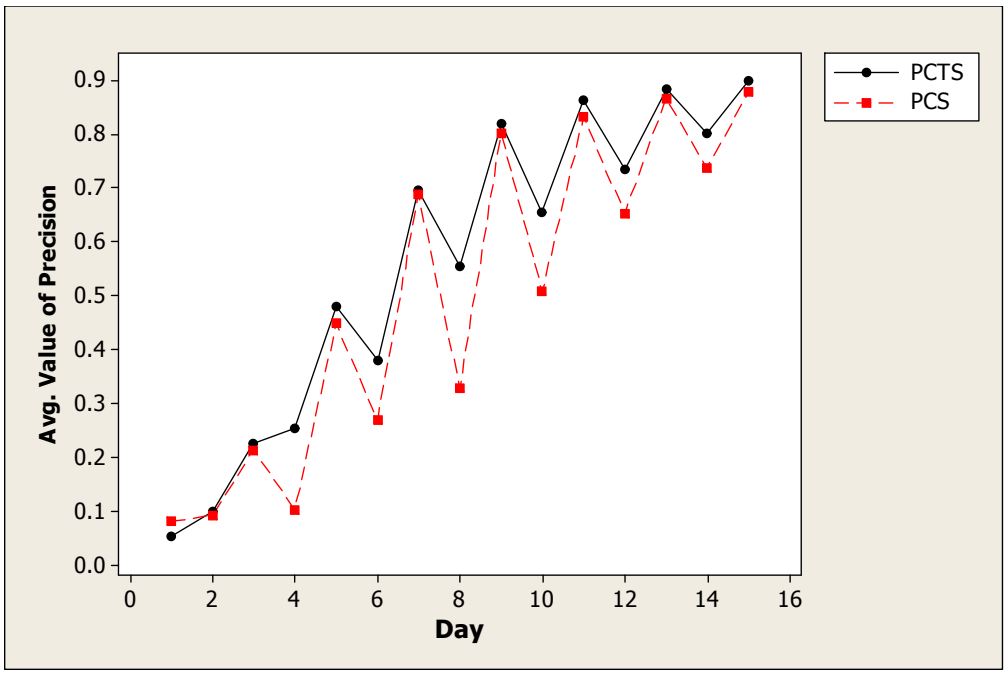


(a) *Cryptosporidium* outbreaks

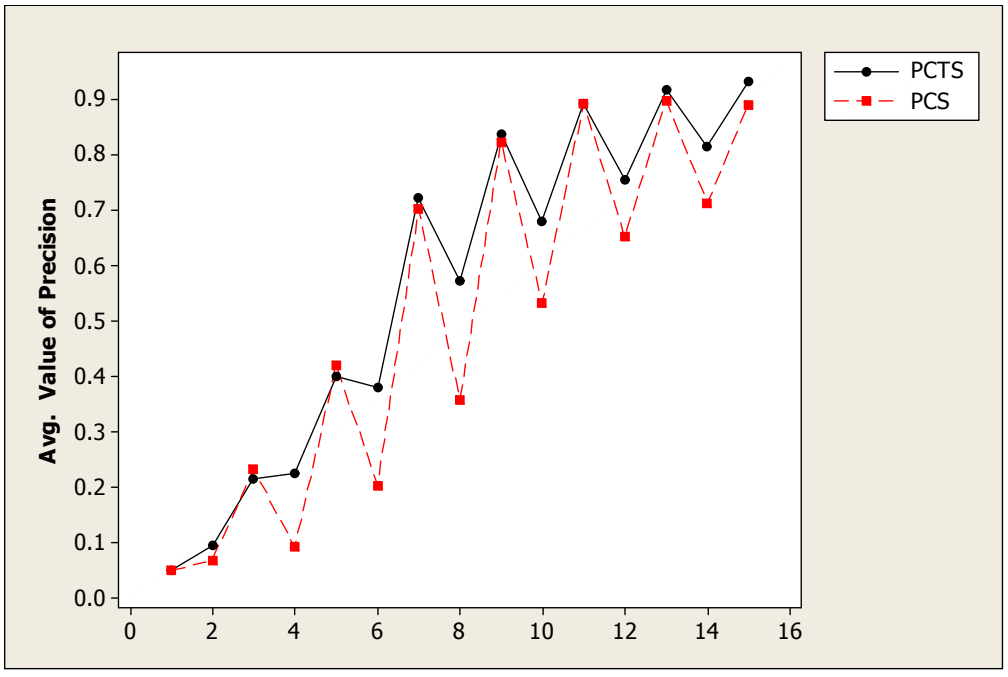


(b) Flu outbreaks

Figure 4.13: The average values of the overlap coefficent for PCTS and PCS.

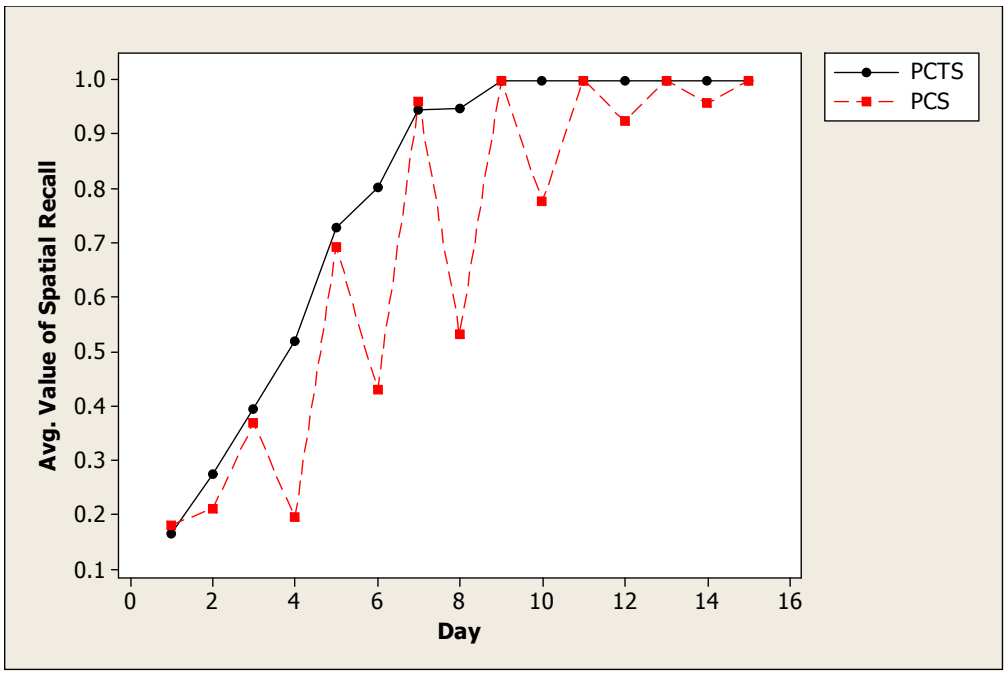


(a) *Cryptosporidium* outbreaks

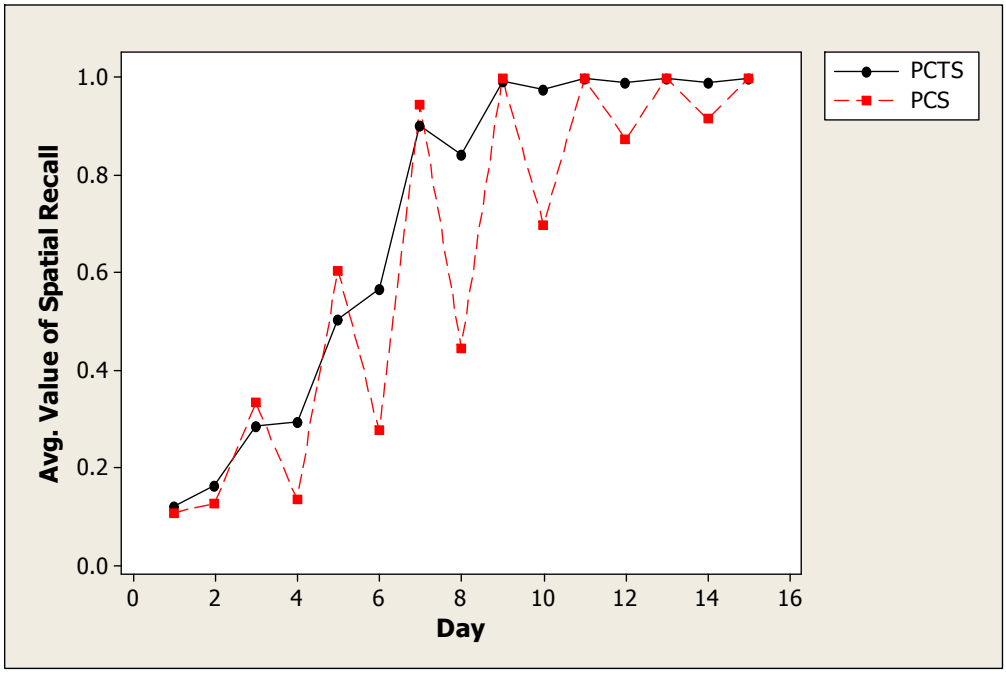


(b) Flu outbreaks

Figure 4.14: The average values of the precision for PCTS and PCS.



(a) *Cryptosporidium* outbreaks



(b) Flu outbreaks

Figure 4.15: The average values of spatial recall for PCTS and PCS.

5.0 ADDITIONAL EXPERIMENTS

In this chapter I show the results of experiments, which do not directly address the hypotheses of this thesis. Rather they compare the extensions of PC (PCS, PCT, and PCTS) to existing state-of-the art outbreak detection systems, and in doing so provide information with which to assess the performance and contributions of these systems.

5.1 EXPERIMENTS COMPARING PCS TO SATSCAN^{TM} AND BSS

The detection power of PCS was compared to the spatial scan statistic (SaTScan^{TM}) (Kulldorff, 1997, 1999), and the Bayesian spatial scan statistic (BSS) (Neil et al., 2005a). I used the SaTScan^{TM} software package, which is available free at http://www.satscan.org/, to implement the spatial scan statistic, the implementation of BSS implemented by Daniel Neill, and my own implementation of PCS. SaTScan^{TM} searched over circular subregions, while BSS and PCS searched over rectangular subregions.

An important question is whether good results can be obtained using a Bayesian network model even if the conditional probability distributions that are in the detection model significantly different from distributions used to generate outbreak data. This question was also investigated in these experiments.

5.1.1 Method

As was done in Section 4.3.1, Allegheny County, Pennsylvania was modeled using a 16×16 grid. Again, both influenza and *Cryptosporidium* outbreaks were simulated in rectangular

subregions of that county. The properties of the simulations were as follows.

1. **Outbreak Severity**: The method for determining outbreak severity was exactly like that discussed in Section 4.3.1. Outbreaks were simulated in which the average daily number of injected ED visits in each cell was equal to $2 \times \sigma_{cell}$.
2. **Daily Increase:** The daily increase was exactly like that discussed in Section 4.3.1. However, only outbreaks that exhibited a linear increase were simulated.
3. **Chief Complaint:** In the same was as described in Section 4.3.1, to determine the chief complaint of each injected case, the chief complaint was generated at random using a probability distribution Q of the chief complaints given the outbreak disease (influenza or *Cryptosporidium*). However, I used values of $N = 1$, 5, 30, and ∞ instead of just $N = 5$, and I analyzed each case separately.
4. **Outbreak Subregions**: The outbreak subregions were determined as discussed in Section 4.3.1. I determined four different subregions of each of the types discussed in Section 4.3.1.

The Kullback-Leibler divergence (Kullback and Leibler, 1951) was used to measure the difference between Q and P for each generated distribution Q. For probability distributions Q and P of the same finite random variable, the **Kullback–Leibler divergence** of Q from P is defined to be

$$D_{KL}(P||Q) = \sum_{i=1}^{n} P(i) \log_2 \frac{P(i)}{Q(i)},$$

where n is the number of alternatives.

Example 5.1. *Suppose* $P(1) = 0.1$, $P(2) = 0.9$, $Q(1) = 0.8$, *and* $Q(2) = 0.2$. *Then*

$$\begin{aligned} D_{KL}(P||Q) &= P(1) \log_2 \frac{P(1)}{Q(1)} + P(2) \log_2 \frac{P(2)}{Q(2)} \\ &= .1 \log_2 \left(\frac{0.1}{0.8}\right) + .9 \log_2 \left(\frac{0.9}{0.2}\right) \\ &= 1.65. \end{aligned}$$

Example 5.2. *Suppose* $P(1) = 0.1$, $P(2) = 0.9$, $Q(1) = 0.2$, *and* $Q(2) = 0.8$. *Then*

$$\begin{aligned} D_{KL}(P||Q) &= P(1)\log_2\frac{P(1)}{Q(1)} + P(2)\log_2\frac{P(2)}{Q(2)} \\ &= .1\log_2\left(\frac{0.1}{0.2}\right) + .9\log_2\left(\frac{0.9}{0.8}\right) \\ &= 0.05. \end{aligned}$$

For each type of outbreak (influenza and *Cryptosporidium*), I developed the number of outbreaks described by the table that follows. Recall that N is our subjective prior sample size. $N = \infty$ denotes that the exact probability distributions in PC were used. The table also shows the average value of the Kullback-Leibler distances of each generated probability distribution from the probability distribution in PC.

N	# Distributions Generated	# Outbreaks per Distribution	Average KL−Dist (flu)	Average KL−Dist (crypto)
∞	1	240	0	0
30	10	60	.24	1.85
5	10	60	1.54	5.86
1	10	60	10.97	22.88

For $N = \infty$, the properties of the 240 outbreaks were determined as follows:

Variable	Values	# Occurrences of Each Value	Total # Occurrences
Duration	30, 40, 50, 60	60	240
Month	$1 - 12$	20	240
Day	$1 - 30$	8	240
Subregion	4 each of types 2 by 1, 2 by 2, and 3 by 2	20	240

For each variable, a list of the 240 occurrences was created. To develop each outbreak, a variable value in the table above was sampled at random without replacement from each list.

For $N = 1$, 5, and 30, the properties of the 60 simulated outbreaks for each of the 10 generated probability distributions were determined as follows:

Variable	Values	# Occurrences of Each Value	Total # Occurrences
Duration	30, 40, 50, 60	15	60
Month	$1-12$	5	60
Day	$1-30$	2	60
Subregion	4 each of types 2 by 1, 2 by 2, and 3 by 2	5	60

Since there were 60 simulated outbreaks for each of the 10 generated distributions, there were a total of 600 simulated outbreaks for each value of N (namely $N = 1$, $N = 5$, and $N = 30$).

I evaluated five methods, namely PCS, two ways of using BSS and two ways of using SaTScan^{TM}. The two ways of using the latter two systems were as follows. The first way looked for a cluster of individuals presenting in the ED with one of the chief complaints that the outbreak disease could cause according to the probability distribution in PC. For example, in the case of influenza, if an individual presented in the ED with any one of the chief complaints, that could be caused by influenza, one was added to the count of observed individuals. There are 12 such chief complaints for *Cryptosporidium* and 20 such chief complaints for influenza. In the second way, I used the probability distribution in PC to determine the three chief complaints that are, according to the criterion developed next, the best indicators of the outbreak disease, and then the systems only looked for clusters of individuals presenting with one of these chief complaints. The criterion for choosing the best indicators is as follow. In the case of influenza, for example, I first assigned a score of 0 to all chief complaints CC such that $P(CC|flu) < \alpha$, where α is a threshold value. I chose $\alpha = 0.002$. In this way, chief complaints that were very unlikely given influenza were not included. Each remaining chief complaint CC was assigned a score as follows:

$$score(CC) = \frac{P(CC|flu)}{P(CC|other\ ED)}.$$

Recall that the value "other" means the individual visited the ED with something other than an outbreak disease. I then ranked the chief complaints by their scores, and chose the top three chief complaints. In this way, these systems were in some way able the take advantage

of the accessed probability distributions in PC. The following tables summarize the results for the top three chief complaints for each outbreak disease:

CC	$P(CC\|flu)$	$P(CC\|other)$	$\frac{P(CC\|flu)}{P(CC\|other)}$
cough	0.3356	0.0248	13.53
fever/chills	0.4122	.0322	12.80
myalgia	0.0095	0.0013	7.308

CC	$P(CC\|Crypto.)$	$P(CC\|other)$	$\frac{P(CC\|Crypto.)}{P(CC\|other)}$
bloody stools	0.03	0.00005	600
sweats	0.1375	0.0003	458.33
diarrhea	0.2643	0.0072	36.71

This table summarizes the inputs to the methods:

Method	Input
PCS	The chief complaint of every individual who visited the ED
BSS Method 1	Count of individuals presenting with any chief complaint caused by injected outbreak disease
BSS Method 2	Count of individuals presenting with one of the top three chief complaints
SaTScan^{TM} Method 1	Count of individuals presenting with any chief complaint caused by injected outbreak disease
SaTScan^{TM} Method 2	Count of individuals presenting with one of the top three chief complaints

Figure 5.1 shows the relationship that would be expected between performance and the two methods of using BSS and SaTScan^{TM}.

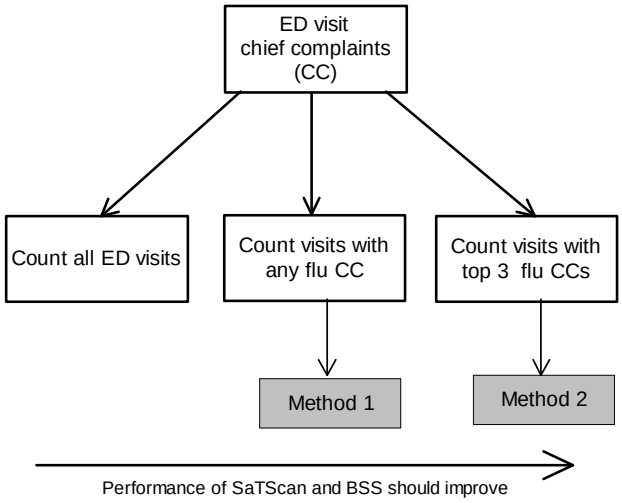


Figure 5.1: As we go to the right the performance of BSS and SaTScan^{TM} are expected to improve.

5.1.2 Results

5.1.2.1 AMOC Curves AMOC curves were used to evaluate the ability of the methods to detect the outbreaks. To create the AMOC curves for SaTScan^{TM}, the likelihood ratio (Equality 2.2) of the most likely subregion was used instead of a posterior probability.

In the evaluation reported in this section, the focus is on the specific detection of *Cryptosporidium* and influenza. For example, Equation 4.2 was used to develop the AMOC curves in the case of the influenza outbreaks. Separate AMOC curves were developed for $N = \infty$, $N = 1$, $N = 5$, and $N = 30$. Recall that for the latter three values of N, 10 probability distributions were generated, and 60 outbreaks were developed using each of the 10 distributions. To obtain a y-value on the AMOC curve, the mean days to detection over all 600 outbreaks was computed.

Figures 5.2 and 5.3 show the AMOC curves comparing the performance of the five methods. We see from these AMOC curves that PCS, BSS Method 2, and SaTScan^{TM} Method 2 all ordinarily performed much better than BSS Method 1 and SaTScan^{TM} Method 1. The

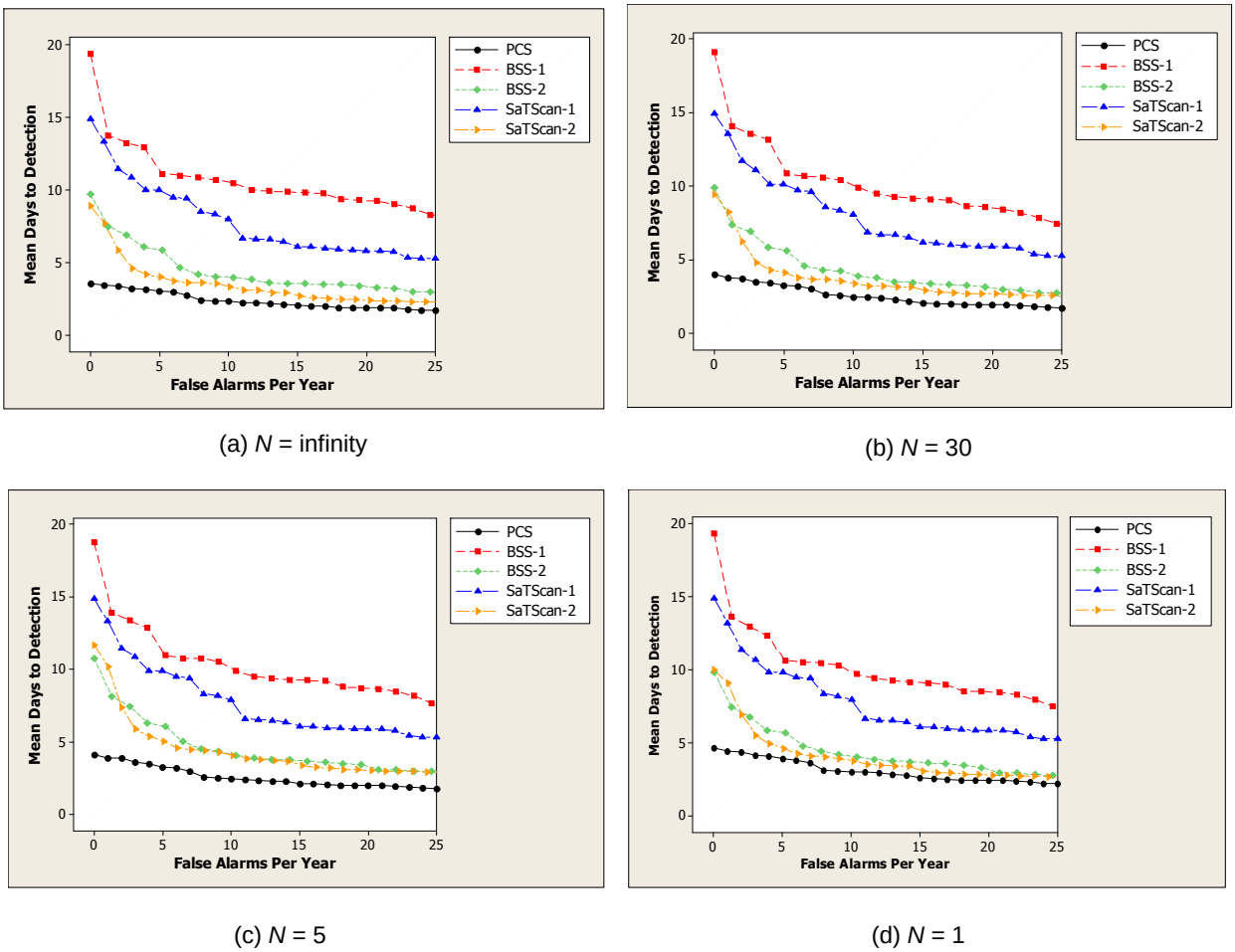
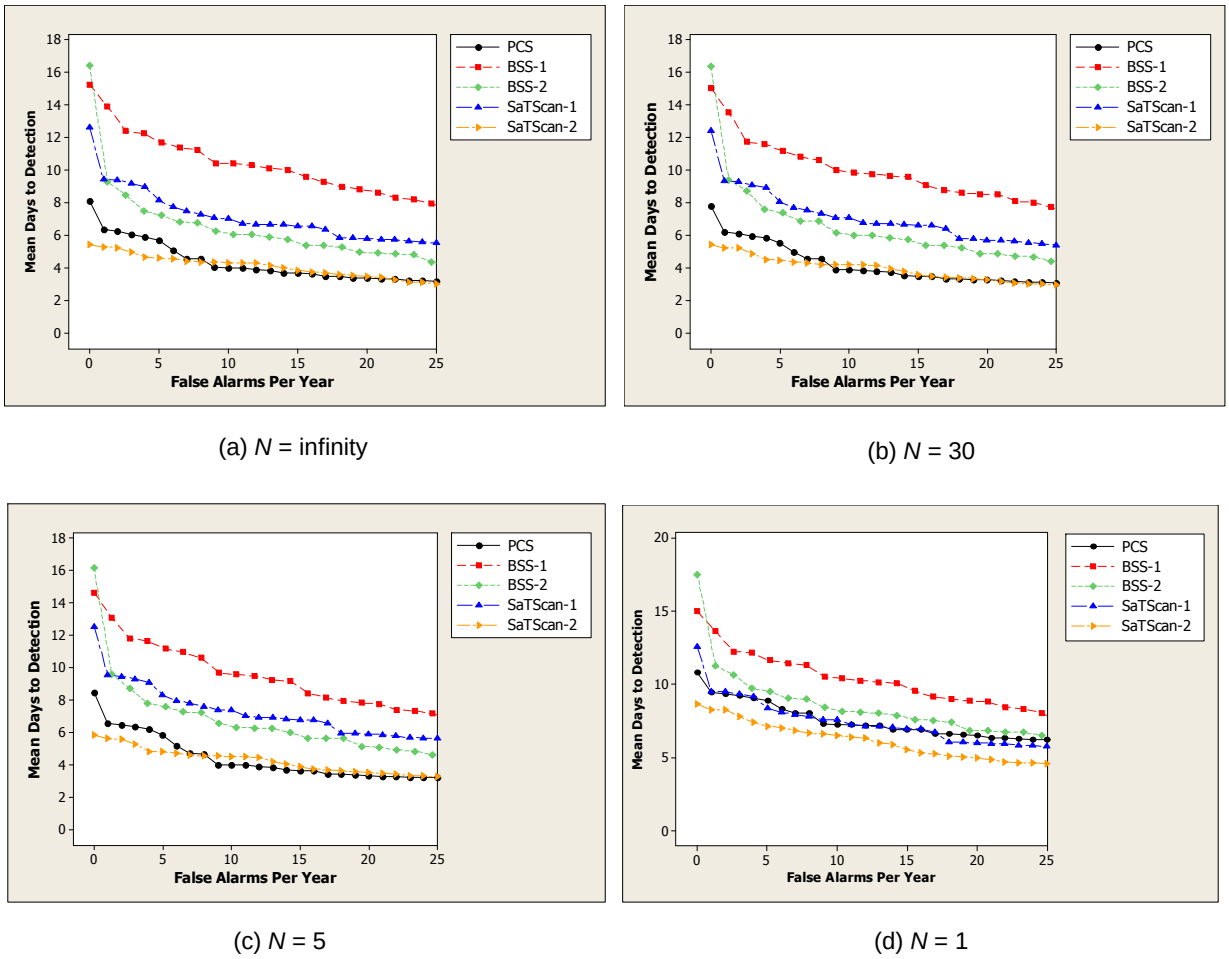


Figure 5.2: AMOC curves comparing the performance of systems when detecting *Cryptosporidium* outbreaks.



(a) N = infinity

(b) N = 30

(c) N = 5

(d) N = 1

Figure 5.3: AMOC curves comparing the performance of systems when detecting influenza outbreaks.

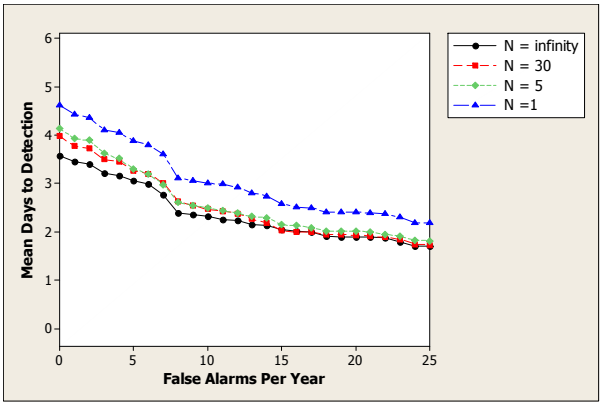
only exception to this is that SaTScan^{TM} Method 1 performed about as well as PCS in the case of influenza outbreaks when $N = 1$. This result substantiates the usefulness of probabilistic information to outbreak detection. Furthermore, the AMOC curves show that in the case of *Cryptosporidium* outbreaks, PCS has better performance than the other methods, and this is particularly true when the number of false alarms per year is fewer than 3. In the case of influenza outbreaks, the performance of SaTScan^{TM} Method 2 exceeds that of PCS, but not by as much as PCS's performance exceeds that of SaTScan^{TM} Method 2 in the case of *Cryptosporidium* outbreaks.

Another important result is that the performance of each of the methods does not degrade very much as the probability distribution used to generate the data increasingly deviates from the one known to the methods. Of course, the performances of BSS Method 1 and SaTScan^{TM} Method 1 would not degrade since these methods do not use probabilistic information. However, even when $N = 1$ the mean day at which PCS detects a *Cryptosporidium* outbreak is 4.63 when the annual false alarm rate is 0, whereas for $N = \infty$ it is 3.56. The corresponding values for influenza outbreak are 10.81 and 8.11 respectively. These results are encouraging, and are consistent with the findings in (Henrion et al., 1996) , which indicated that diagnosis using Bayesian networks is often insensitive to imprecision in probabilities. In the case of outbreak detection, it seems that perhaps, as long as we identify some of the most likely chief complaints given an outbreak disease, we can obtain good detection performance even if imprecision is high.

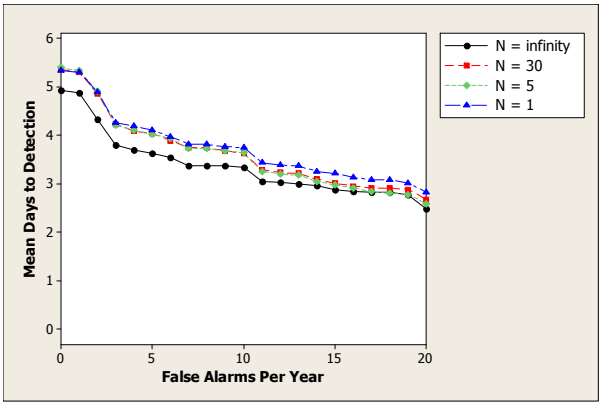
Notice that SaTScan^{TM} Method 2 tended to out-perform BSS Method 2. The outbreaks all occurred in small subregions, and BSS should show better performance relative to SaTScan^{TM} when the outbreak subregions are larger. SaTScan^{TM} looks at circles whose maximum radius is such that the window never exceeds more than 50% of the population at risk. Furthermore, SaTScan^{TM} uses population as its baseline, and considers how the counts in subregion S and the counts in subregion $G - S$ compare relative to the population baselines in those subregions. Therefore, if the subpopulation in outbreak subregion S exceeded 50% of the population, the detection performance of SaTScan^{TM} would degrade. This degradation would increase as the size of S increased until finally there would be no detection capability when $S = G$. On the other hand, BSS looks at all rectangles, and uses

the past 28 days of data to obtain values for the mean and the variance of the disease rate in each subregion. When conditioning on the presence of an outbreak in subregion S, BSS multiplies the computed mean for S by a factor m, which is distributed uniformly between 1 and 3. So if $G - S$ is small or even null, an outbreak in subregion S could still be detected because of the increased conditional probability of the data given that there is an outbreak in S.

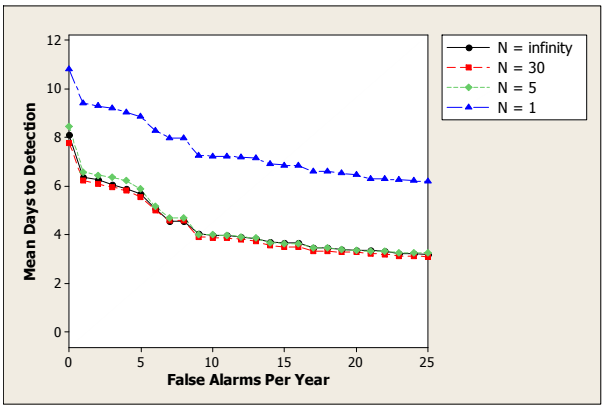
Recall that PCS detects 12 different types of outbreaks. So it not only reports the probability of an influenza or a *Cryptosporidium* outbreak, but also the overall probability of a non-specific outbreak. Figure 5.4 contains AMOC curves showing PCS's performance regarding the detection of a non-specific outbreak. In that figure, this performance is compared to the performance regarding the detection of the specific outbreak which was injected. For example, Equality 4.2 was used to produce the AMOC curve in Figure 5.4 (c), and Equality 4.1 was used to produce the AMOC curve in Figure 5.4 (d). In the case of *Cryptosporidium* outbreaks, PCS can detect a non-specific outbreak a little worse than it can specifically detect a *Cryptosporidium* outbreak. In the case of influenza outbreaks, for most values of N the performance when detecting a non-specific outbreak is about the same as that when detecting an influenza outbreak, and when $N = 1$ it actually detects a non-specific outbreak much better than it detects an influenza outbreak. The result for $N = 5$, 30, and ∞ may be due to the fact that influenza outbreaks are difficult to detect because influenza has symptoms such as cough and fever/chills, which are not uncommon when no outbreak is occurring. Since influenza outbreaks are difficult to detect, it seems reasonable that it would be no easier to detect an influenza outbreak than a non-specific outbreak. The result for $N = 1$ seems reasonable for the same reason. That is, perhaps, when the conditional probability distributions of the chief complaints given influenza in the detection model are significantly different from the conditional probability distributions used to generate the outbreak data, it may be quite difficult to detect an influenza outbreak. However, since there is still a substantial increase in ED visits during the outbreak, it may not be any more difficult to detect a non-specific outbreak. Note that *Cryptosporidium*, on the other hand, has very distinct symptoms.



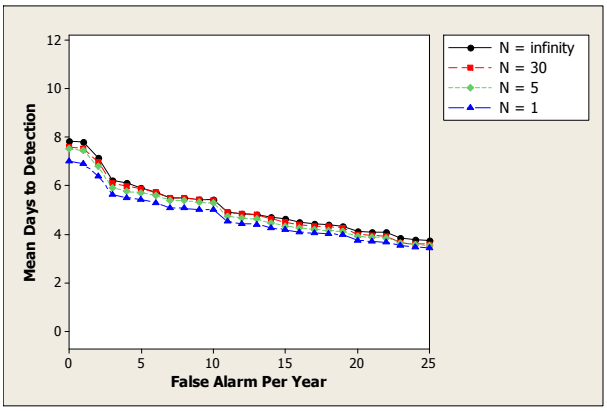
(a) Mean days to detecting a *Cryptosporidium* outbreak when a *Cryptosporidium* outbreak is injected



(b) Mean days to detecting any outbreak when a *Cryptosporidium* outbreak is injected



(c) Mean days to detecting a flu outbreak when a flu outbreak is injected



(d) Mean days to detecting any outbreak when a flu outbreak is injected

Figure 5.4: AMOC curves comparing PCS's ability to detect any outbreak (non-disease specific) to its ability to detect the specifically simulated outbreak disease.

FAR	$P(\mu_{Sc} > \mu_{Pc})$	$P(\mu_{Bc} > \mu_{Pc})$	$P(\mu_{Sf} > \mu_{Pf})$	$P(\mu_{Bf} > \mu_{Pf})$
0	1	1	**0**	1
5	1	1	**0.025**	0.896
10	1	0.9998	**0.0075**	0.971
15	0.9230	0.9999	**0.0075**	0.936
20	0.9985	0.9998	**0.0004**	0.806

Table 5.1: At various false alarm rates (FAR), the posterior probabilitiy that PCS has a smaller mean day to detection than another method when specifically detecting the simulated outbreak disease, when N=1.

5.1.2.2 Significance Testing of Detection Power Recall that PCS not only reports the posterior probability of an influenza or a *Cryptosporidium* outbreak, but also the overall probability of an outbreak. In the statistical significance tests, I compared both the performance of PCS concerning the detection of the outbreak disease and its performance concerning the detection of a non-specific outbreak to the performances of BSS Method 2 and SaTScan$^{\text{TM}}$ Method 2. I did the comparison for $N = 1$ because, based on the AMOC curves in Figures 5.2 and 5.3, that seems to be the situation in which PCS performed poorest relative to the other two. Tables 5.1 and 5.2 show the results. In those tables the following notation was used to refer to each system:

S_c: SaTScan$^{\text{TM}}$ Method 2 detecting *Cryptosporidium* outbreaks.

B_c: BSS Method 2 detecting *Cryptosporidium* outbreaks.

P_c: PCS detecting *Cryptosporidium* outbreaks.

S_f: SaTScan$^{\text{TM}}$ Method 2 detecting influenza outbreaks.

B_f: BSS Method 2 detecting influenza outbreaks.

P_f: PCS detecting influenza outbreaks.

P_{nc}: PCS detecting non-specific outbreak during *Cryptosporidium* outbreaks.

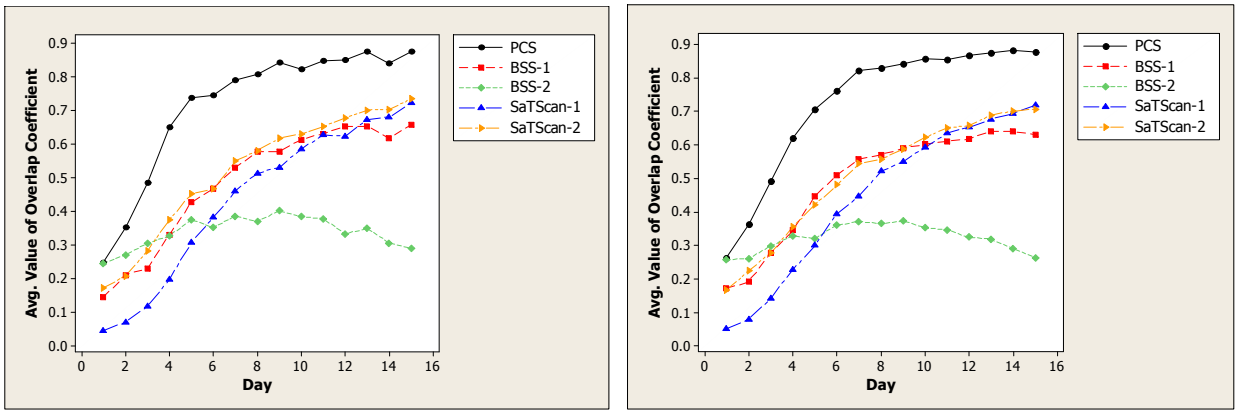
P_{nf}: PCS detecting non-specific during influenza outbreaks.

FAR	$P(\mu_{Sc} > \mu_{Pnc})$	$P(\mu_{Bc} > \mu_{Pnc})$	$P(\mu_{Sf} > \mu_{Pnf})$	$P(\mu_{Bf} > \mu_{Pnf})$
0	1	1	1	1
5	0.9999	1	1	1
10	0.9437	0.8338	1	1
15	0.618	0.9243	1	1
20	0.639	0.9678	1	1

Table 5.2: At various false alarm rates (FAR), the posterior probabilitiy that PCS has a smaller mean day to detection than another method when detecting a non-specific outbreak, when N=1.

We see from Table 5.1 that, when PCS is detecting the injected outbreak, in most cases it is most probable that PCS has a smaller mean time to detection. The exceptions to this are highlighted in that table. In these highlighted cases, we are comparing PCS's ability to detect an influenza outbreak to the performance of SaTScan$^{\text{TM}}$ Method 2. It has already been noted that when $N = 1$, PCS is much better at detecting a non-specific outbreak than it is at detecting an influenza outbreak. We see from Table 5.2 that, when PCS is detecting a non-specific outbreak, in all cases it is most probable that PCS has a smaller mean time to detection than the other systems.

5.1.2.3 Subregion Detection The subregion S_j that maximized $P(Data|SUB = S_j)$ was considered to be the subregion detected by PCS and the Bayesian spatial scan statistic. The subregion that maximized the Poisson spatial scan statistic was considered to be the subregion detected by SaTScan$^{\text{TM}}$. Figures 5.5 and 5.6 show the average values of the overlap coefficients on each day of the outbreaks. Notice that the variable *Day* (on the x-axis) never exceeds 15 even though the simulation lengths were between 30 and 60 days. The reason is that the simulations were only run for 15 days even though their theoretical lengths were greater than 15. In all cases the performance of PCS was significantly better than that of the other methods. In the case of *Cryptosporidium* outbreaks, PCS performed



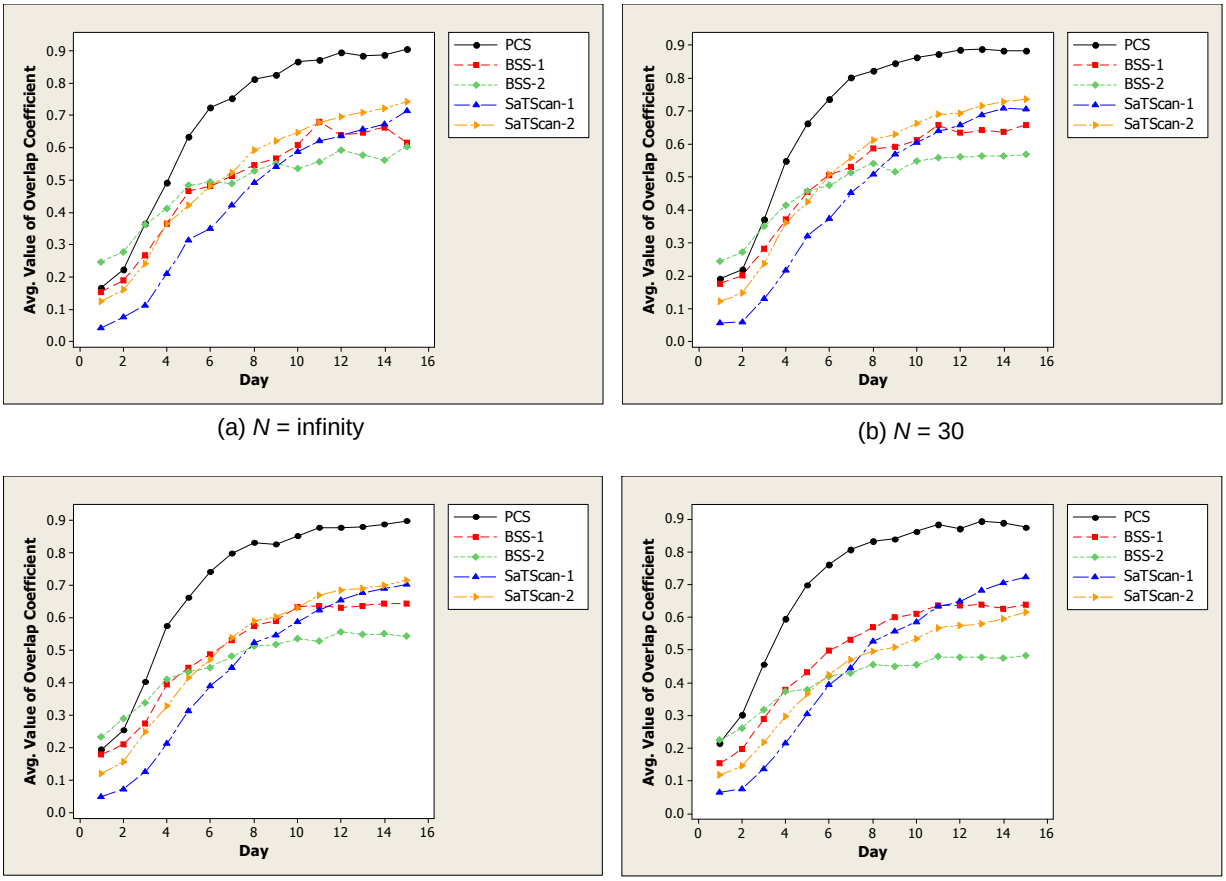
(a) N = infinity

(b) N = 30

(c) N = 5

(d) N = 1

Figure 5.5: The average values of the overlap coefficient for *Cryptosporidium* outbreaks.



(a) N = infinity

(b) N = 30

(c) N = 5

(d) N = 1

Figure 5.6: The average values of the overlap coefficient for influenza outbreaks.

substantially better than SaTScan^{TM} Method 2 at both outbreak detection and subregion detection. However, in the case of influenza outbreaks, PCS performed about the same as SaTScan^{TM} Method 2 at outbreak detection (once the false alarm rate is greater than four), but it substantially outperformed SaTScan^{TM} Method 2 at subregion detection. As discussed in Section 5.1.2, PCS was better at detecting a non-specific outbreak than it was at specifically detecting an influenza outbreak, and its performance when detecting a non-specific outbreak was significantly better than the performance of SaTScan^{TM}. This is consistent with its superior subregion detection performance (which is not tied to which type of outbreak disease is detected).

Furthermore, PCS's performance did not degrade as the probability distributions used to generate the data increasingly deviated from the simulation model's distributions (i.e., as N went from ∞ to 1). These results are expected since the number of cases injected into the outbreak subregion does not depend on N. That is, if we changed the type of chief complaints injected but not their number, we may detect a different outbreak, but we would not change the subregion most likely to contain the outbreak. In the same way, the performance of Methods 1 of SaTScan^{TM} and BSS should not deteriorate as we decrease N, but the performance of Methods 2 should deteriorate because they only consider a subset of chief complaints. Figures 5.5 and 5.6 are consistent with these expectations.

5.1.3 Summary

I compared the performance of PCS, SaTScan^{TM}, and BSS. PCS outperformed the other two methods both in terms of outbreak detection and subregion detection. This was the case even when SaTScan^{TM} and BSS were able to take advantage of the probabilistic information in PCS. These results lend support to the conjecture that, in the case of spatial event surveillance, we may be able to obtain better results by modeling the relationships among the events of interest and the observable events using a Bayesian network, rather than using summary statistics.

Perhaps more importantly, the performance of PCS was very robust relative to the probability distribution generating the data. It is an open question whether we can obtain

acceptable results using a Bayesian network if the probability distributions in the network, which are often obtained from limited data and/or subjective judgement, do not closely reflect reality. The results shown here indicate that in the domain of disease outbreak detection this seems to the case.

Finally, I found that PCS can detect the presence of a non-specific outbreak almost as well (indeed, sometimes better) as it detects a specific outbreak.

Recall that SaTScan^{TM} searched over circles, while BSS and PCS searched over rectangles, and our injected subregions were rectangles. So it seems BSS and PCS had an advantage. However, this advantage is not as great as one might think for the following reason. Even though rectangular subregions were chosen, the injections actually occurred in zip codes whose centroids were in the cells in the subregions. So the actual injected subregions were not rectangles.

Since some of the subregions were 1 cell by 2 cells and others were 2 cells by 3 cells, it may have been better if SaTScan^{TM} searched over ellipses. In some cases, the SaTScan^{TM} software package does allow us to search over elliptical subregions. However, the information that we were able to provide consisted of zip codes, which is information about longitude and latitude. The SaTScan^{TM} software package does not allow searching over elliptical subregions in this case. Regardless, since our rectangles were not very elongated and since the injected subregions were actually the zip codes whose centroids were in the rectangles, it seems that searching over circular subregions instead of elliptical subregions should not significantly change the results.

5.2 EXPERIMENTS COMPARING PCTS TO MULTIVARIATE, TEMPORAL SATSCAN^{TM}

In these experiments I compared PCTS to the multivariate, temporal version of the spatial scan statistic (Kulldorff, 2004), which I designate as SaTScan^{TM}-MT. My purpose was to determine how PCTS fares relatively to a state-of-the-art multivariate, spatio-temporal cluster detection system.

5.2.1 Method

The method was identical to that described in Section 4.3.1. Indeed, the same set of simulations were used.

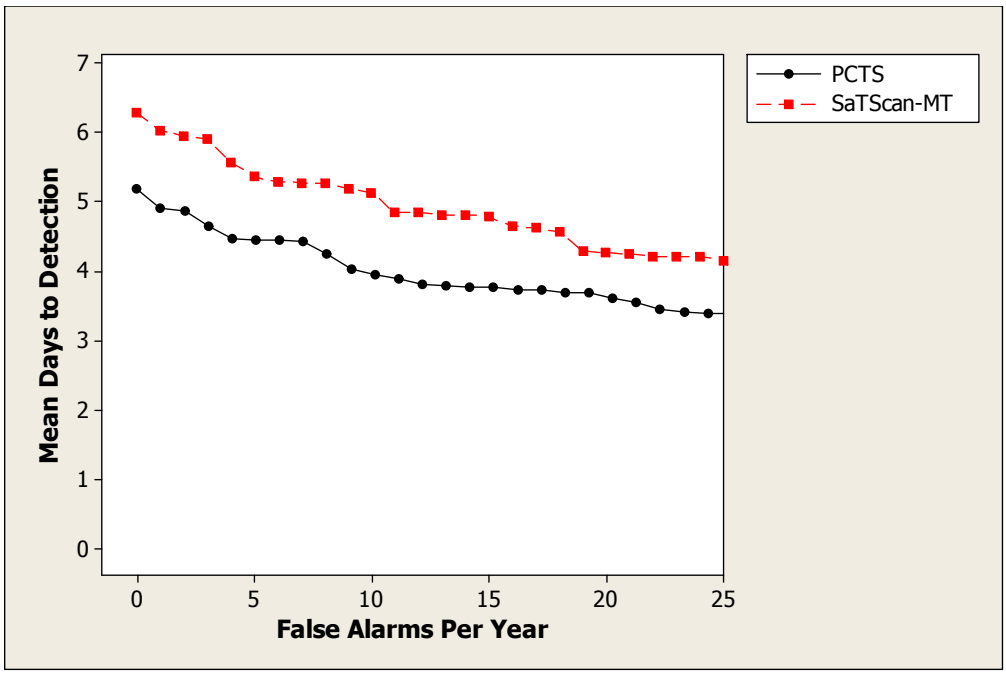
5.2.2 Results

5.2.2.1 AMOC Curves Figure 5.7 shows AMOC curves comparing the detection performance of PCTS and SaTScan^{TM}-MT. In the case of *Cryptosporidium* outbreaks, PCTS performed noticeably better than SaTScan^{TM}-MT. In the case of influenza outbreaks, PCTS perform better for very small false alarm rates and for large false alarm rates, but worse for rates in the middle. Overall, PCTS's performance appears to be better.

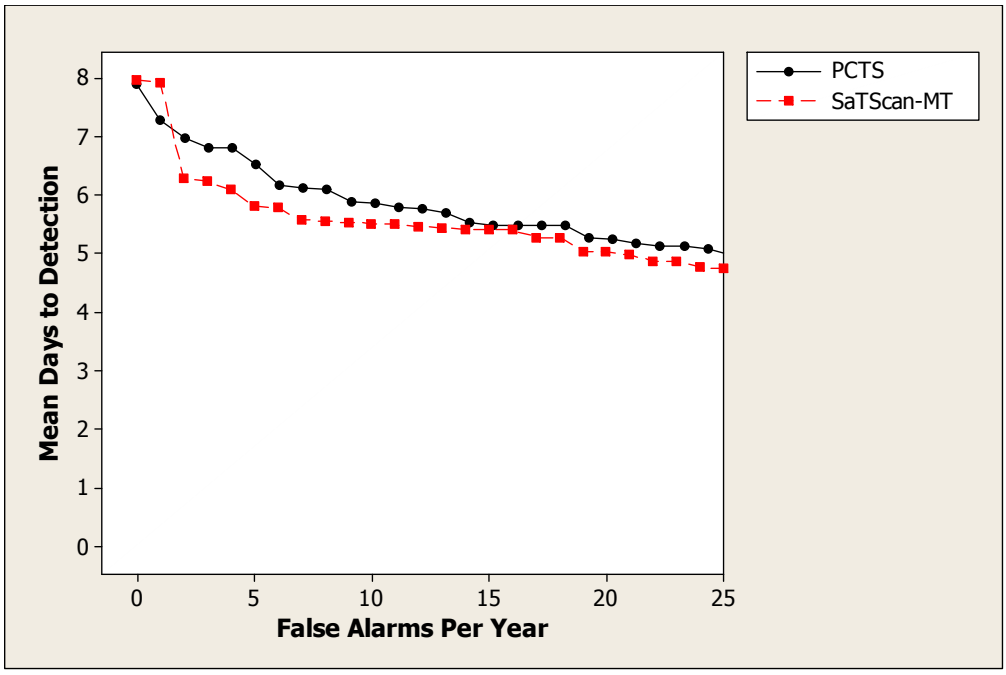
5.2.2.2 AMOC-M Curves Figure 5.8 shows AMOC-M curves comparing the detection performance of PCTS and SaTScan^{TM}-MT. In the case of *Cryptosporidium* outbreaks, PCTS performed noticeably better than SaTScan^{TM}-MT, while in the case of influenza outbreaks SaTScan^{TM}-MT performed much better than PCTS. For completeness I note that, in the case of influenza outbreaks, PCTS did perform better than SaTScan^{TM}-MT when the false alarm rate is one. Overall, these results do not seem to indicate better performance for either system.

5.2.2.3 Subregion Detection The subregion S that maximized $P(Data|SUB = S)$ was considered to be the subregion detected by PCTS, and the subregion that maximized the Poisson spatial scan statistic was considered to be the subregion detected by SaTScan^{TM}-MT.

Figures 5.9 shows the average values of the overlap coefficient for PCTS and SaTScan^{TM}-MT. PCTS outperformed SaTScan^{TM} -MT for both types of outbreaks.

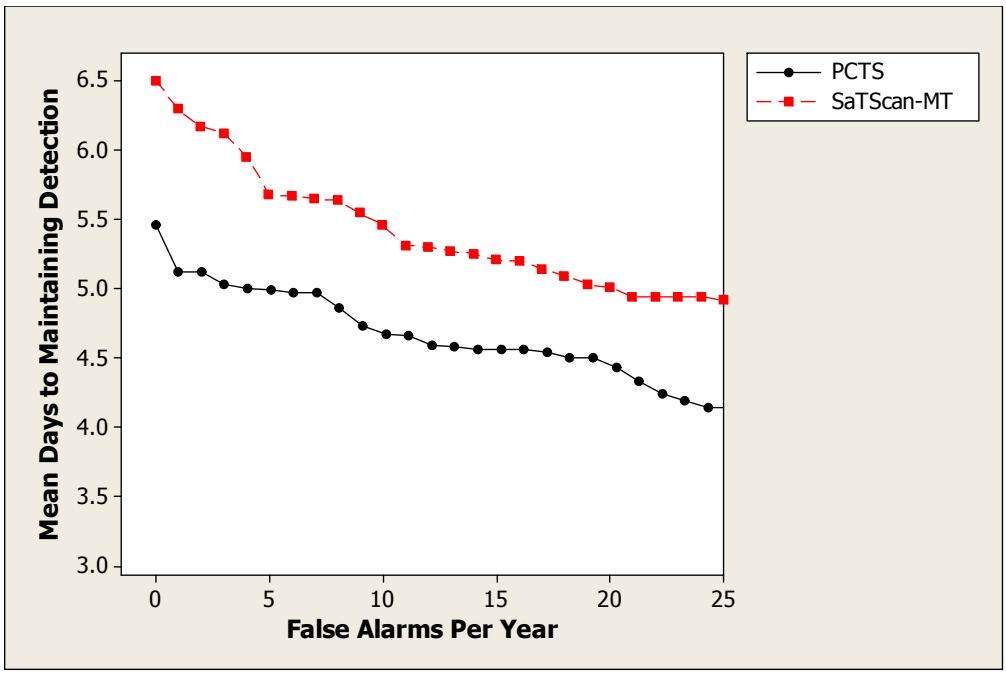


(a) *Cryptosporidium* outbreaks

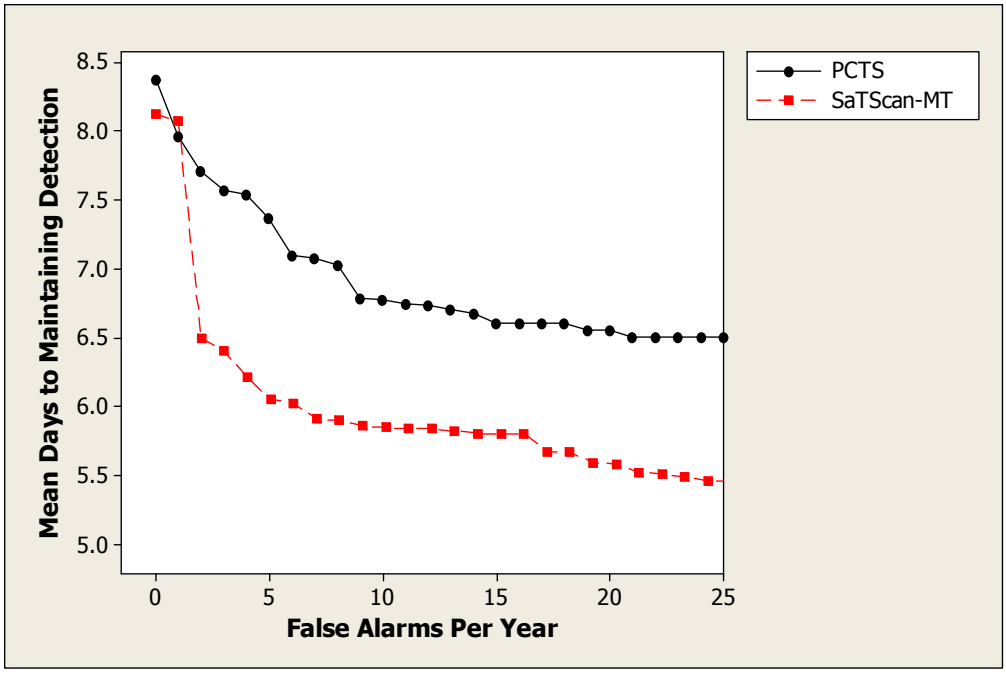


(b) Flu outbreaks

Figure 5.7: AMOC curves comparing the detection performance of PCTS and SaTScan^{TM}-MT.

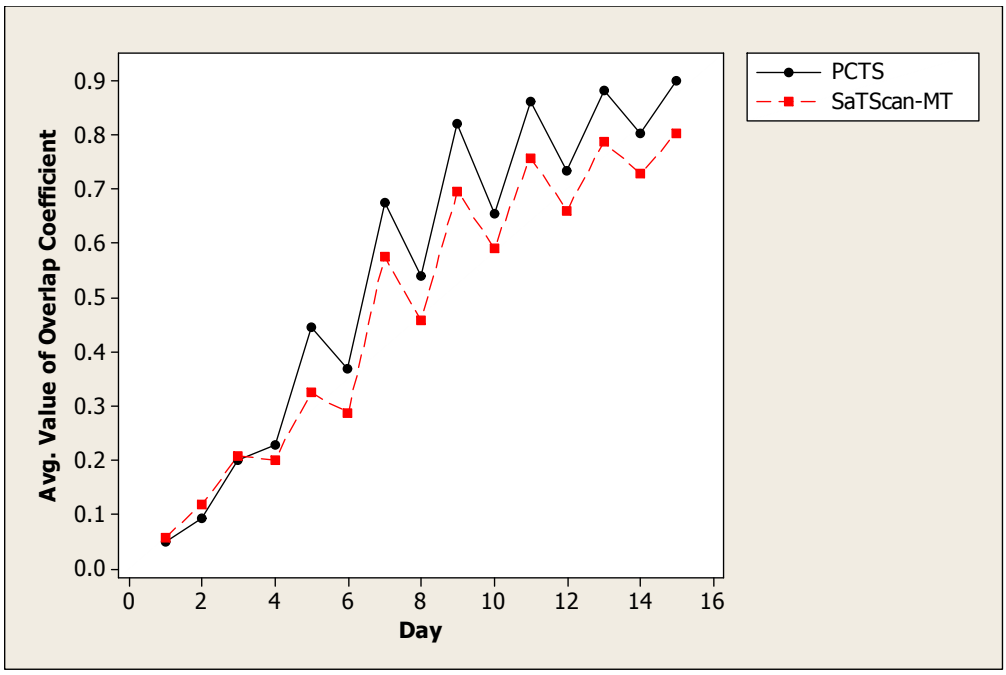


(a) *Cryptosporidium* outbreaks

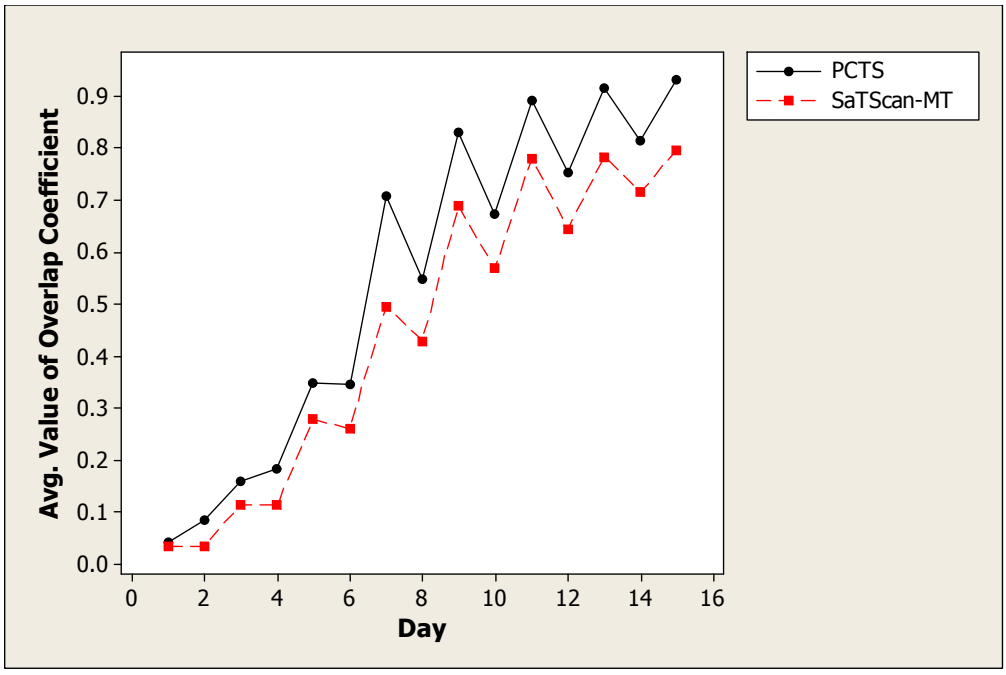


(b) Flu outbreaks

Figure 5.8: AMOC-M curves comparing the detection maintenance performance of PCTS and SaTScan^{TM}-MT.



(a) *Cryptosporidium* outbreaks



(b) Flu outbreaks

Figure 5.9: The average values of the overlap coefficient for PCTS and SaTScan^{TM}-MT.

5.3 EXPERIMENTS COMPARING PCT TO CUSUM

In these experiments I compared PCT to CUSUM (Bos and Fetherston, 1992) to determine how PCT performs relative to a classic temporal outbreak detection system.

5.3.1 Method

The method was identical to that described in Section 4.3.1. Indeed, the same set of simulations were used.

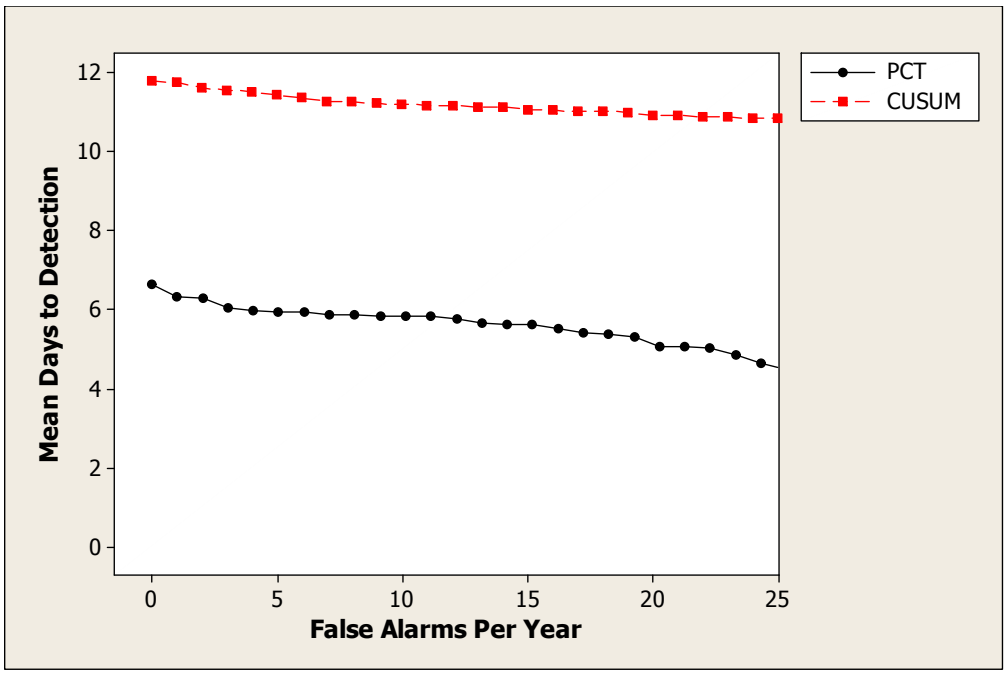
5.3.2 Results

Figure 5.10 shows AMOC curves comparing the detection performance of PCT and CUSUM. Notice that CUSUM in Figures 5.10 (b) stops at a false alarm rate of 10. CUSUM returns a number r where $r \geq 0$. Higher numbers are more indicative of an outbreak. There were only 10 days in the background on which this number exceeded 0. So if we signaled an outbreak whenever $r > 0$, we would have a false alarm rate equal to 10. The next possible value at which we could signal an outbreak is whenever $r \geq 0$. However, then we would have a false alarm rate of 365 with a mean day of detection equal to 0.

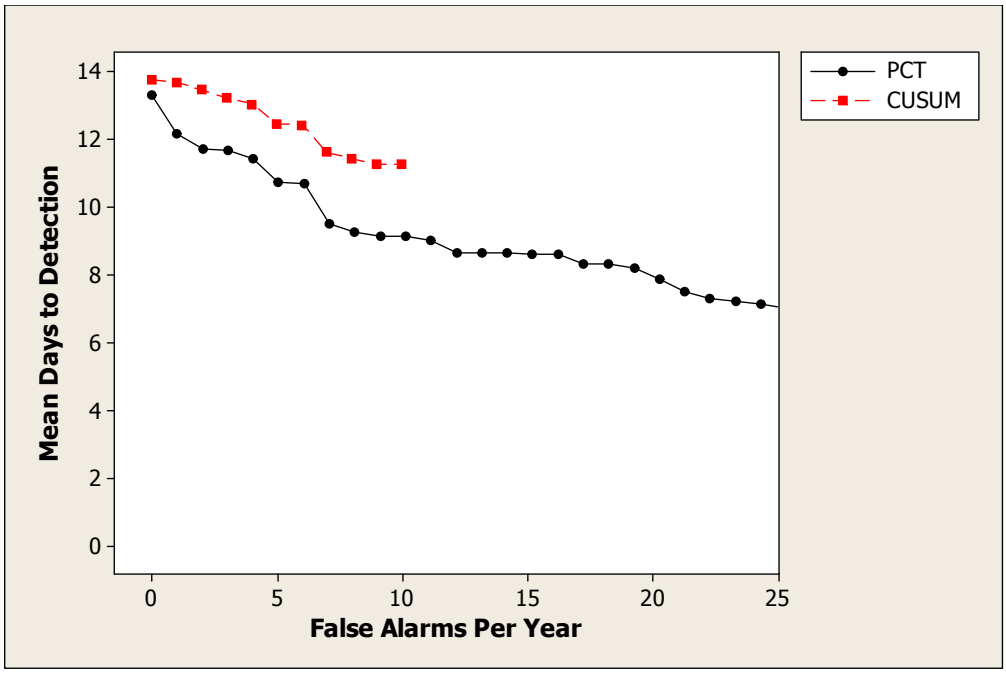
PCT performed substantially better than CUSUM in the case of the *Cryptosporidium* outbreaks, and noticeably better in the case of the influenza outbreaks.

5.4 EXPERIMENTS CONCERNING BINARY VERSIONS OF PCS, PCT, AND PCTS

Binary PCS (B-PCS) is a version of PCS in which the observable variable is binary. The observable variables in PCS are multinomial because their values can be any one of 54 chief complaints with which an individual presents to the ED. In B-PCS our data for each individual consists only of whether the individual arrived at the ED with one of the three primary chief complaints of an outbreak disease being evaluated. The primary chief complaints are



(a) *Cryptosporidium* outbreaks



(b) Flu outbreaks

Figure 5.10: AMOC curves comparing the detection performance of PCT and CUSUM.

the ones specified in Section 5.1.1. So B-PCS has the same input as BSS Method 2 and SaTScan^{TM} Method 2 had in that section. B-PCT and B-PCTS are likewise binary versions of PCT and PCTS, respectively.

I compared the performance of each system to its binary version to determine whether there is any advantage in multinomial modeling in this domain.

5.4.1 Method

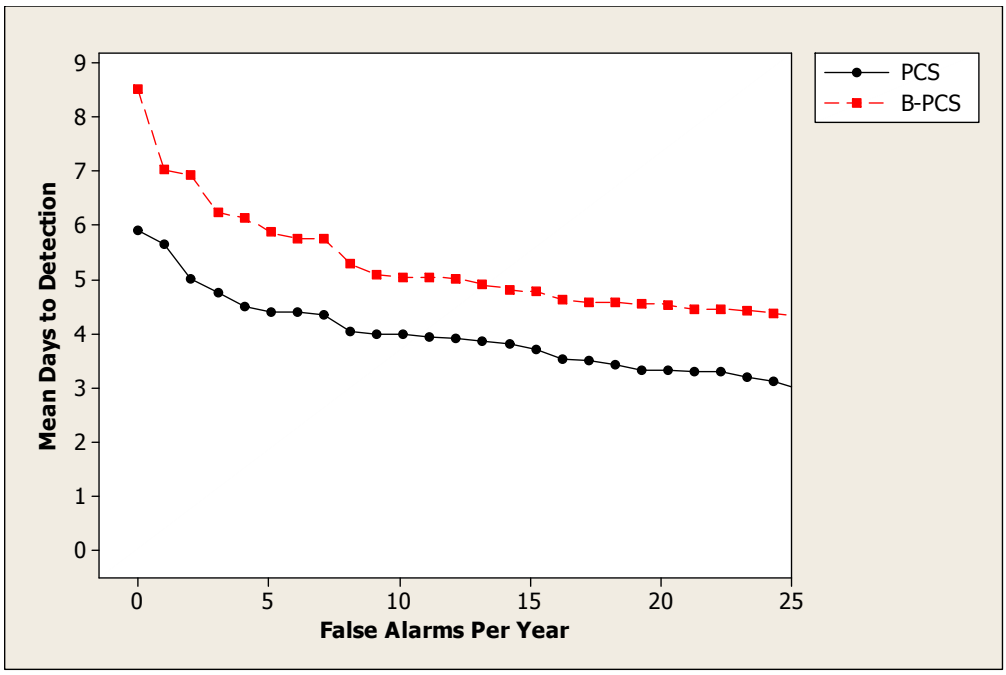
The method was identical to that described in Section 4.3.1. Indeed, the same set of simulations were used.

5.4.2 Results

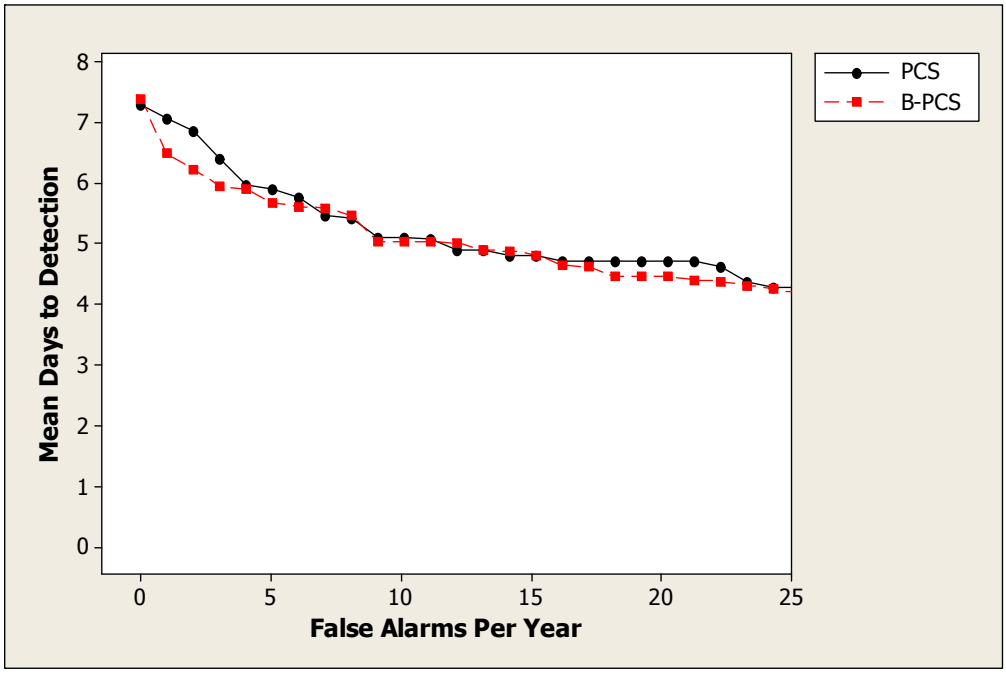
Figure 5.11, 5.12, and 5.13 show AMOC curves comparing the detection performance of the systems to their binary counterparts. In the case of *Cryptosporidium* outbreaks, PCS, PCT, and PCTS all performed much better than B-PCS, B-PCT, and B-PCTS respectively. In the case of influenza outbreaks, PCS and PCT performed about the same as their binary counterparts, whereas PCTS performed worse than B-PCTS. The results for the *Cryptosporidium* outbreaks indicate that multinomial modeling can be useful. As to the results for influenza, recall that the AMOC curves in Figure 5.3 indicate that SaTScan^{TM} Method 2 seems to be a little better at detecting influenza outbreaks than PCS. Like B-PCT, SaTScan^{TM} Method 2 looks only at the three primary chief complaints. Perhaps in the case of influenza we obtain better detection performance by only considering the three primary chief complaints.

5.5 EXPERIMENTS CONCERNING OUTBREAKS EMERGING IN SPACE

In this set of experiments, I further compared PCTS and SaTScan^{TM}-MT by evaluating how well they detected outbreaks that were emerging in both time and space.

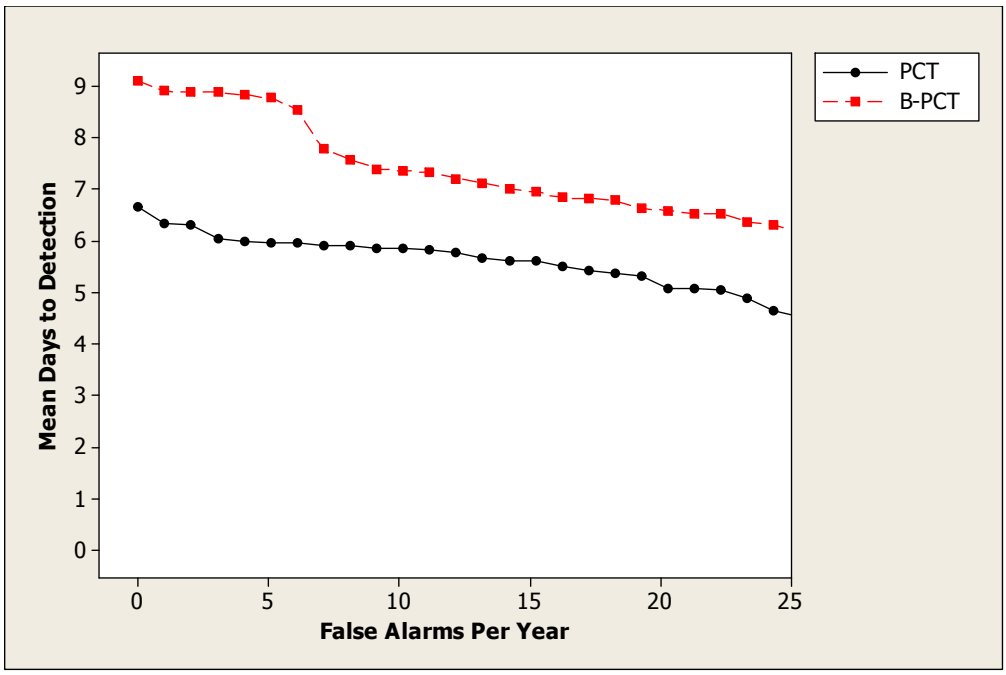


(a) *Cryptosporidium* outbreaks

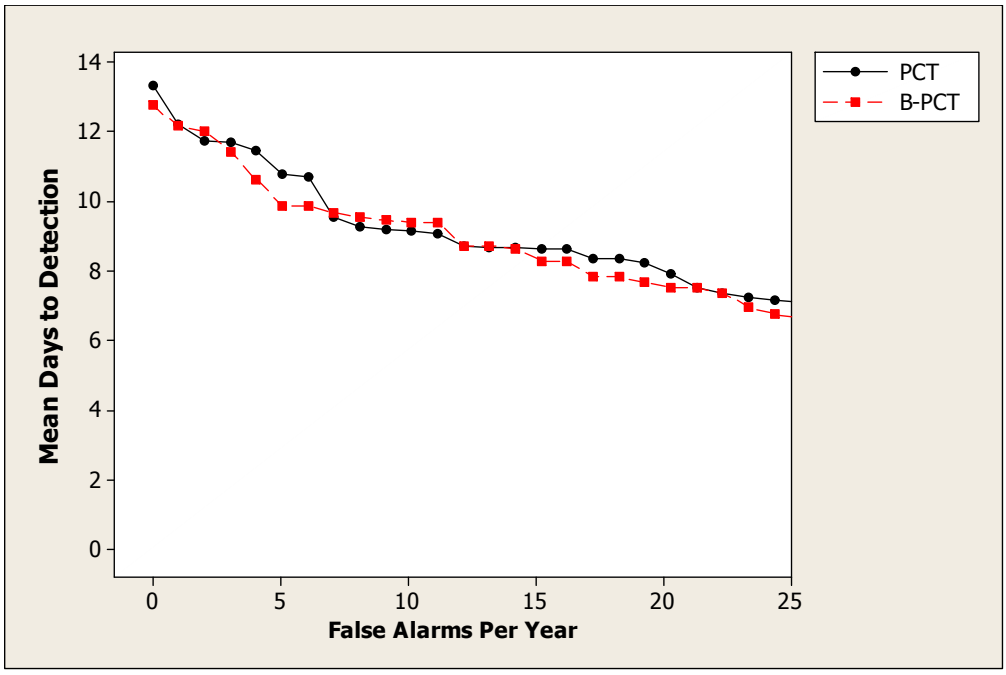


(b) Flu outbreaks

Figure 5.11: AMOC curves comparing the detection performance of PCS and B-PCS.

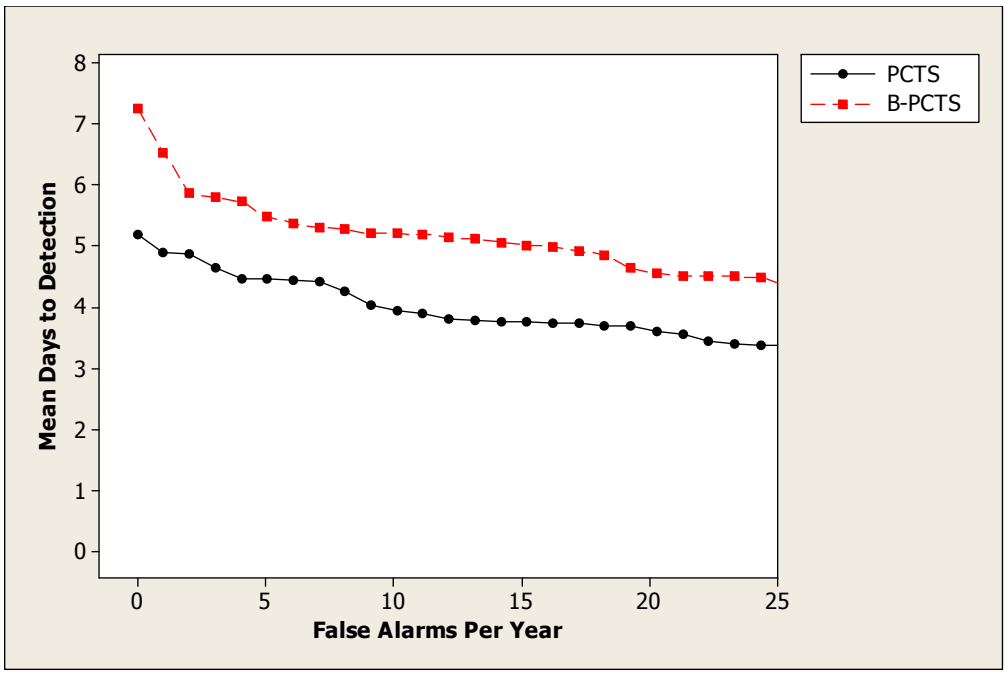


(a) *Cryptosporidium* outbreaks

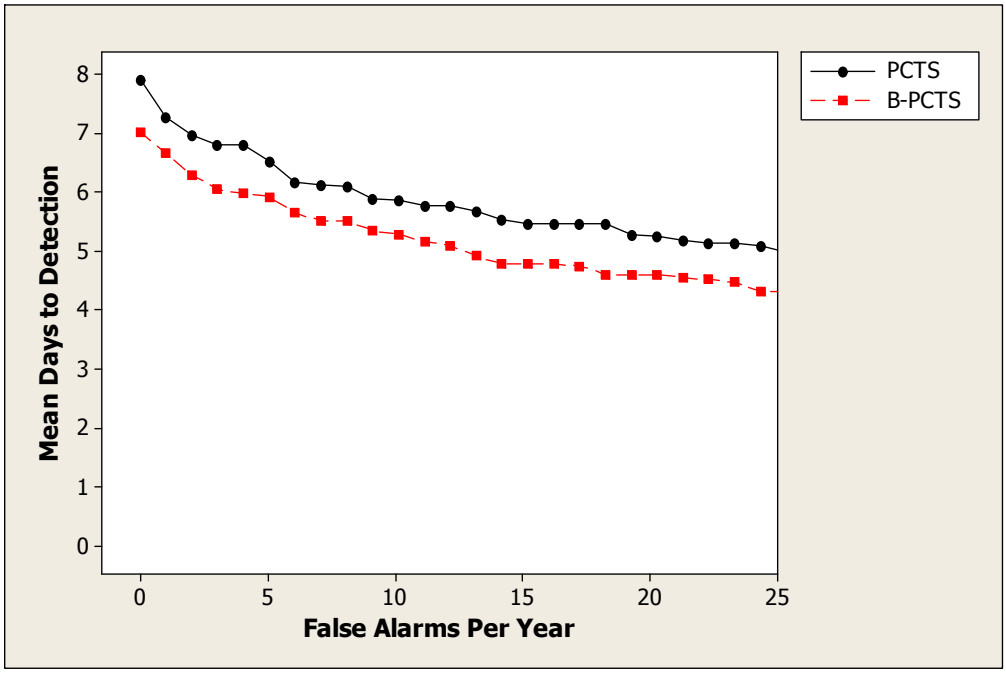


(b) Flu outbreaks

Figure 5.12: AMOC curves comparing the detection performance of PCT and B-PCT.



(a) *Cryptosporidium* outbreaks



(b) Flu outbreaks

Figure 5.13: AMOC curves comparing the detection performance of PCTS and B-PCTS.

5.5.1 Method

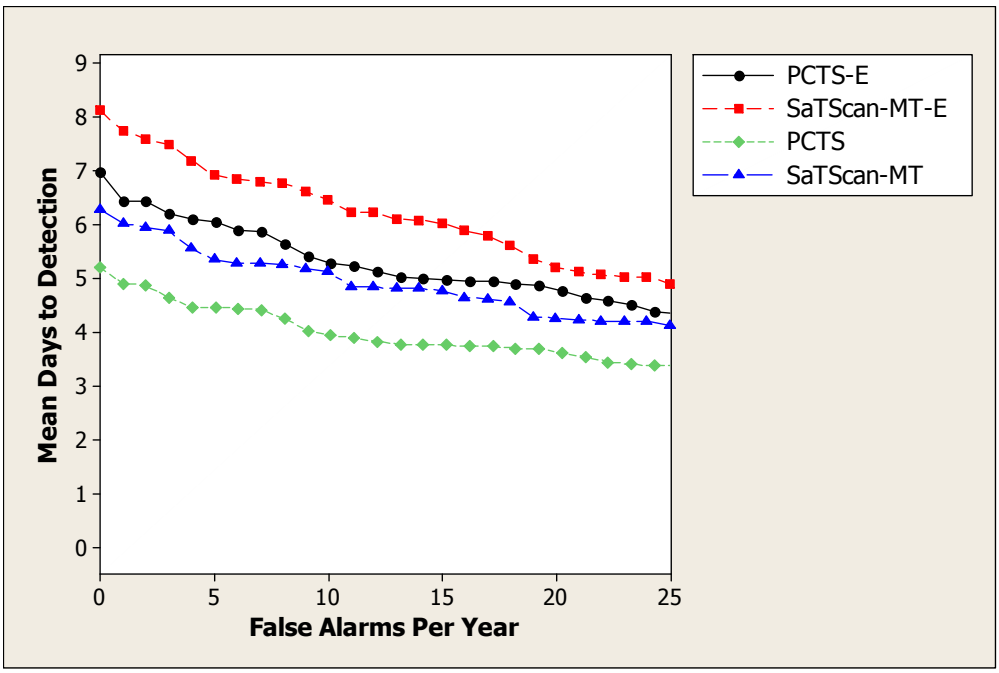
The method was exactly like that described in 4.3.1 except that the simulated outbreaks were also made to emerge in space. I modeled time emergence by injecting cases into 1 cell on days 1 and 2 of the outbreak, into 2 cells on days 3 and 4, into 3 cells on days 5 and 6, and so until all cells in the injected subregion were receiving injections. I injected cases into 1 cell by randomly selecting one of the cells in the injected subregion, I injected cases into 2 cells by randomly selecting the second cell from all the cells that touched the first cell, and I followed the same procedure when I injected cases into 3 or more cells in the injected subregion.

5.5.2 Results

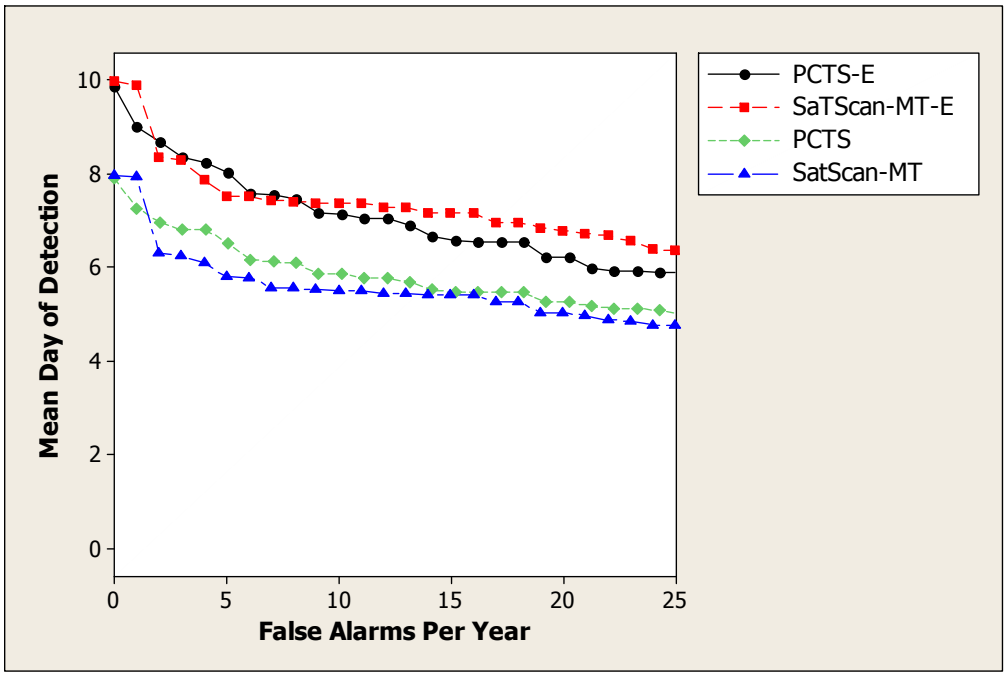
5.5.2.1 AMOC Curves Figure 5.14 shows AMOC curves comparing the performance of PCTS and SaTScan^{TM}-MT when detecting the emerging outbreaks. The curves labeled with the letter "E" are the ones concerning the emerging outbreaks. The other curves show the previous results concerning outbreaks that are not emerging. We would expect the results to be worse for emerging outbreaks because initially not as many cells have cases injected. The results are consistent with this expectation.

In the case of the *Cryptosporidium* outbreaks, PCTS performed substantially better for both emerging and non-emerging outbreaks. In the case of influenza outbreaks, PCTS and SaTScan^{TM}-MT performed similarly.

5.5.2.2 Subregion Detection Figure 5.15 shows the average values of the overlap coefficient for PCTS and SaTScan^{TM}-MT on each day of the emerging outbreaks. They also show the results for the non-emerging outbreaks. PCTS outperformed SaTScan^{TM}-MT in all cases. That is, regardless of whether the outbreak is influenza or *Cryptosporidium*, and regardless of whether it is emerging or non-emerging, PCTS performed better.

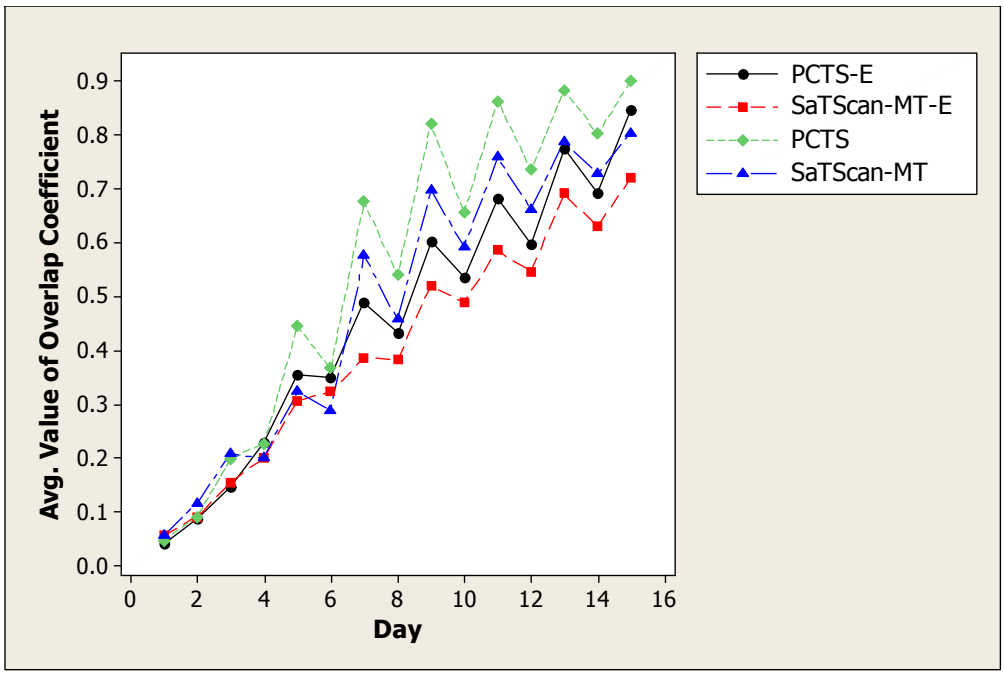


(a) *Cryptosporidium* outbreaks

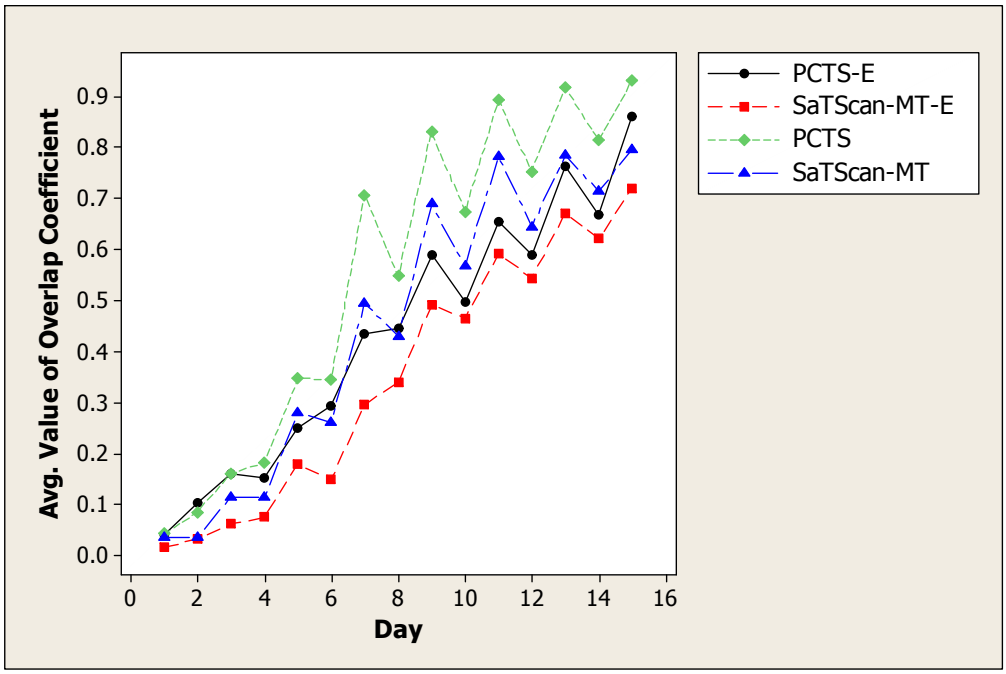


(b) Flu outbreaks

Figure 5.14: AMOC curves comparing several systems.



(a) *Cryptosporidium* outbreaks



(b) Flu outbreaks

Figure 5.15: The average values of the overlap coefficent for several systems.

5.6 EXPERIMENTS CONCERNING ONE-STEP OUTBREAKS

In this set of experiments, I investigated how well the temporal model (PCT) performed relative to its non-temporal counterpart (PC) in a situation which seems difficult for the temporal model. That is, I simulated outbreaks in which the number of new outbreak cases does not gradually increase each day, but instead the number of new cases jumps to a value one day into the outbreak, and then each day after that the number of new cases stays fixed at that value.

5.6.1 Method

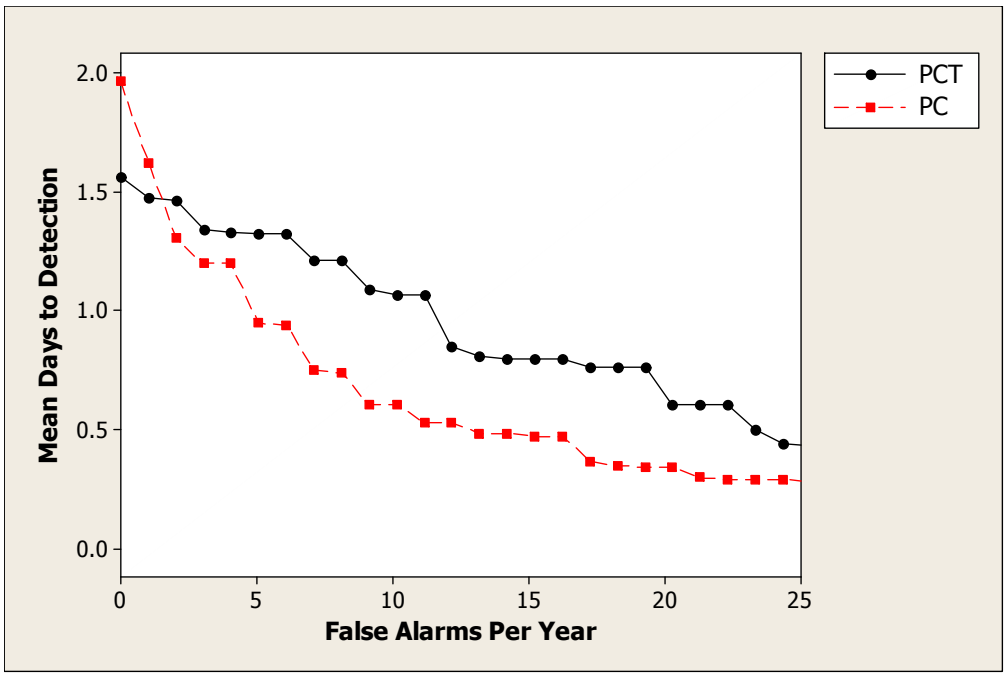
The method was exactly like that described in 4.3.1 except that I simulated outbreaks whose injection curve was a single **step function**. That is, I injected the same number of cases on every day of the outbreak. Outbreaks with the following levels of severity were simulated:

Severity Level	Avg. Daily # Injected ED Visits
1	$1.5\sigma_{cell}$
2	$2\sigma_{cell}$
3	$2.5\sigma_{cell}$
4	$3\sigma_{cell}$

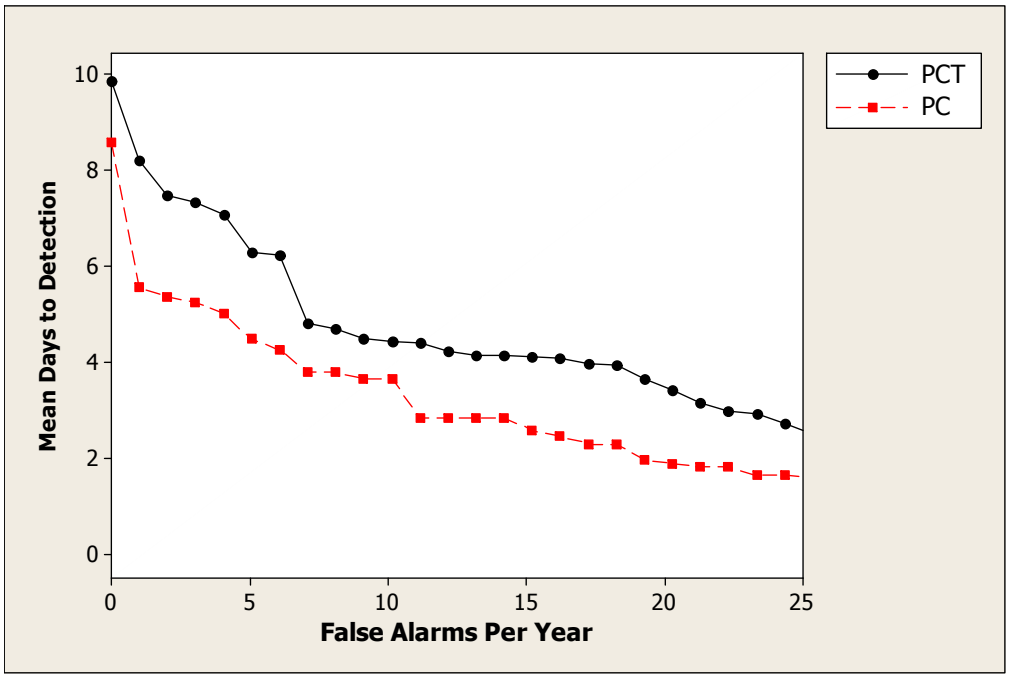
If, for example, a severity level 2 outbreak was being simulated and $\sigma_{cell} = 4$, then the daily number of injections in the cell would be $2 \times 4 = 8$.

For each outbreak disease, I performed 120 simulations. The properties of the 120 simulated outbreaks were determined as follows:

Variable	Values	# Occurrences of Each Value	Total # Occurrences
Prob. Dist.	$1, 2, 3, 4, 5, 6, 7, 8, 9, 10$	12	120
Month	$1 - 12$	10	120
Day	$1 - 30$	4	120
Severity	$1, 2, 3, 4$	30	120
Subregion	4 each of types 2 by 1, 2 by 2, and 3 by 2	10	120



(a) *Cryptosporidium* outbreaks



(b) Flu outbreaks

Figure 5.16: AMOC curves comparing PCT and PC when detecting outbreaks that increase in one step.

For each variable, a list of the 120 occurrences was created. To develop each outbreak, a variable value in the table above was sampled at random without replacement from each list.

5.6.2 Results

AMOC curves comparing PC and PCT appear in Figure 5.16. As expected, PCT performed substantially worse than PC when detecting both types of outbreaks. However, the performance is not as bad as we might have thought. First, at a false alarm rate of 0, PCT actually performed better than PC when detecting *Cryptosporidium* outbreaks. Second, the difference in the detection means is less than two days for influenza outbreaks, and it is less than 0.5 days for *Cryptosporidium* outbreaks. Recall that the value $T = 5$ was used for the parameter T in PCT, which means PCT analyzed up to 4 days of data besides the current day. Therefore, on the first day of the outbreak, PCT was considering up to four days of data with no injected ED visits and one day of data with injected ED visits. On the other hand, on the first day of the outbreak PC was considering precisely one day of data with injected ED visits. So PC was looking only at outbreak data from day one of the outbreak, while PCT was not looking only at outbreak data until day five of the outbreak.

5.7 FURTHER COMPARISONS OF PCT TO PC

Recall from Section 4.3.3 that PCT performed about the same as PC, regarding outbreak detection, when we looked at results aggregated over all the outbreaks. However, perhaps PCT and PC perform differently for certain subsets of outbreaks. First, I show AMOC curves comparing their results for two cases in which it seems their performance may differ. Then I perform logistic regression on four variables to learn which variables may affect the relative performances of PCT and PC.

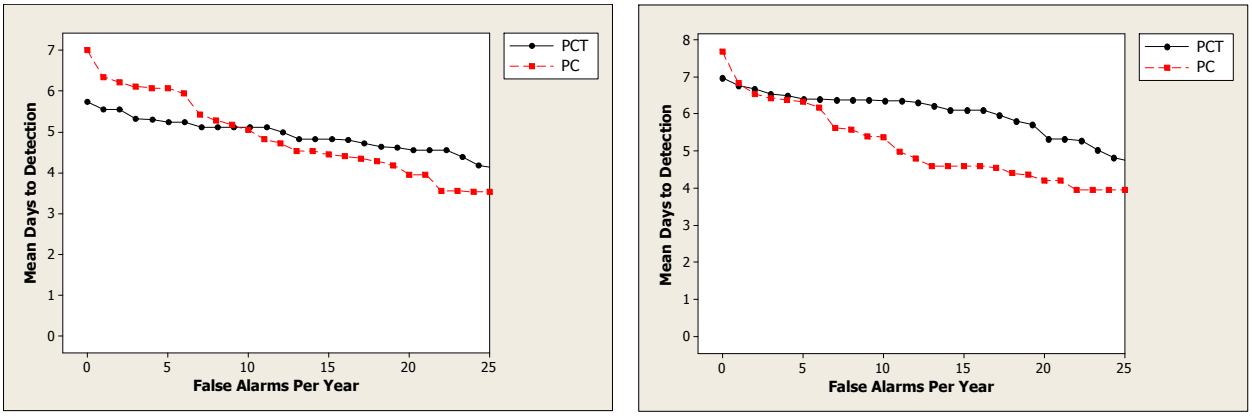
5.7.1 AMOC Curves Comparing Impact

5.7.1.1 Impact of Epidemic Curve Function The results in Section 5.6.2 indicate that PC outperforms PCT in outbreak detection in the case of an outbreak with an extreme sudden onset (step function). Based on these results, it seems that the performance of PCT relative to PC would improve as the onset becomes less sudden. If so PCT should exhibit better performance for linear-increasing outbreaks than for quadratic-increasing outbreaks. Similarly, its performance for quadratic-increasing outbreaks should be better than its performance for cubic-increasing outbreaks. I investigated the relative performances of PCT and PC separately for linear, quadratic, and cubic-increasing outbreaks using the simulations described in Section 4.3.1. The results appear in Figures 5.17 and 5.18.

These results mildly substantiate my conjecture. In the case of *Cryptosporidium* outbreaks, PCT performs noticeably better than PC when the false alarm rate is less than 10 in the case of linear-increasing outbreaks, and this is not true for the other types of increases. In the case of influenza outbreaks, PCT performs better than PC when the false alarm rate is 0 in the case of linear-increasing outbreaks, and this also is not true for the other types of increases.

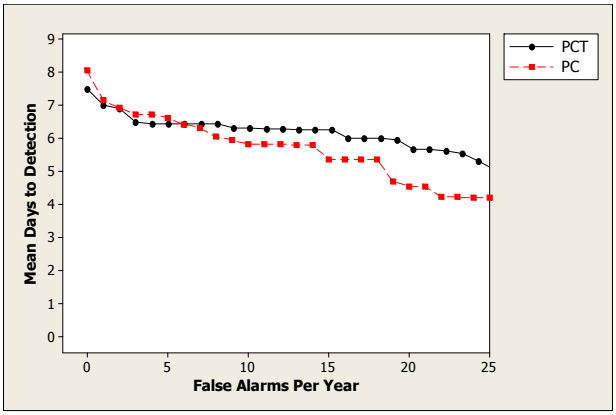
Note that another explanation for PCT's performance being better for linear-increasing outbreaks is that the PCT model assumes that the outbreak exhibits a linear increase.

5.7.1.2 Impact of Fluctuation on Performance Recall that in half the simulations, on even numbered days I made the number of new cases 25% of the previous day's number, and in the other half I made it 50% of the previous day's number. It seems that PCT may exhibit better comparative performance to PC when I use the value 50% relative to when I use the value 25% because there will be higher counts on the even numbered days. In general, it seems that PCT's comparative performance to PC should improve as the daily fluctuation decreases. Figure 5.19 shows the results separately for fluctuation values of 25% and 50%. In the case of both *Cryptosporidium* outbreaks and influenza outbreaks, PCT exhibits better comparative performance to PC when the fluctuation is less (fluctuation value of 50%) than when the fluctuation is greater (fluctuation value of 25%). These results

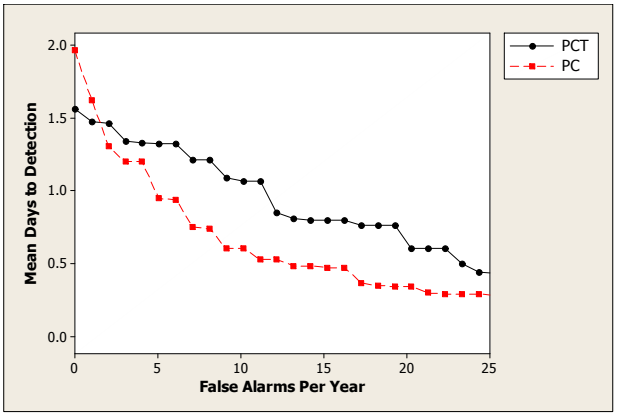


(a) *Cryptosporidium* (linear)

(b) *Cryptosporidium* (quadratic)

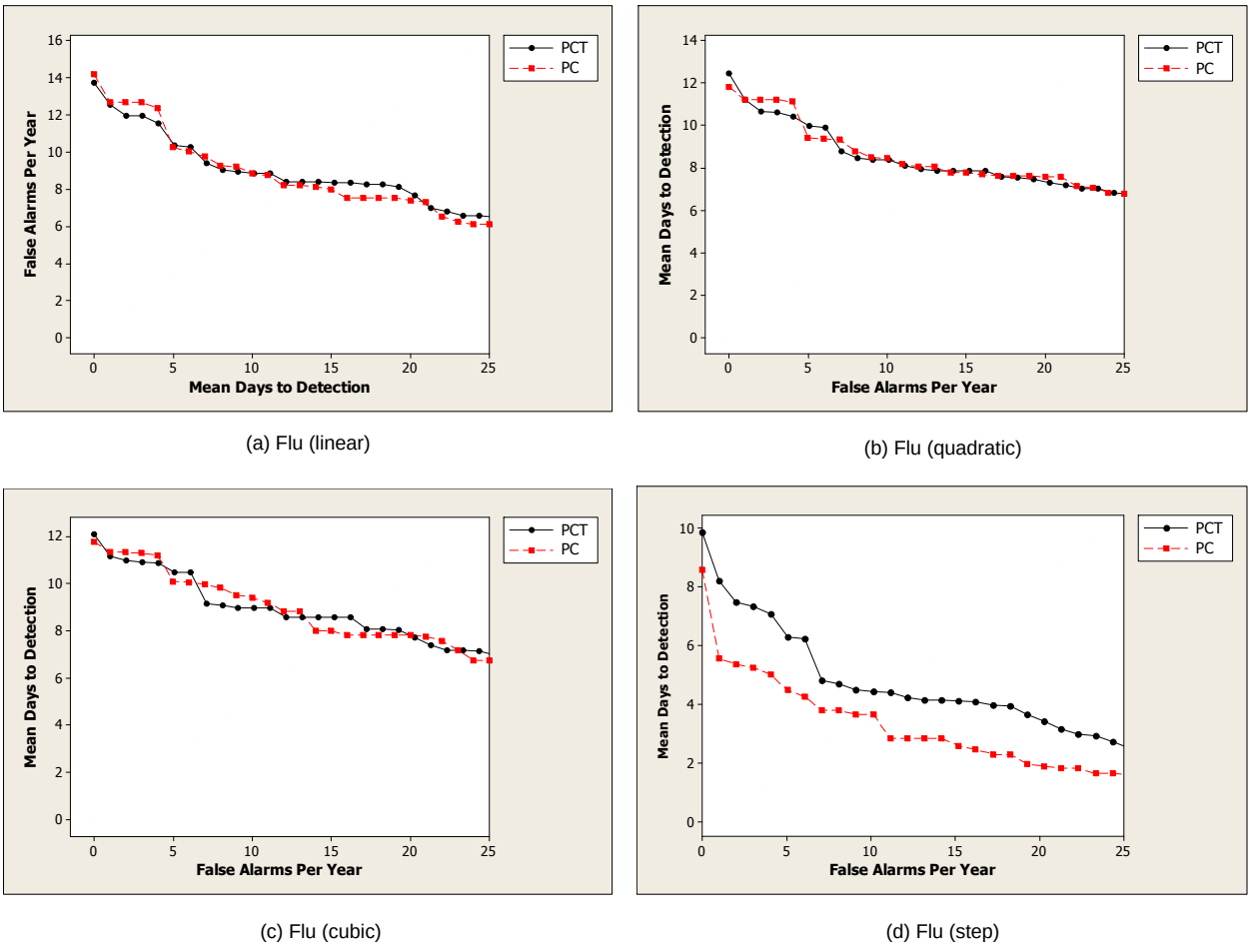


(c) *Cryptosporidium* (cubic)



(d) *Cryptosporidium* (step)

Figure 5.17: The relative performances of PCT and PC for linear-increasing, quadratic-increasing, cubic-increasing, and step-function *Cryptosporidium* outbreaks.



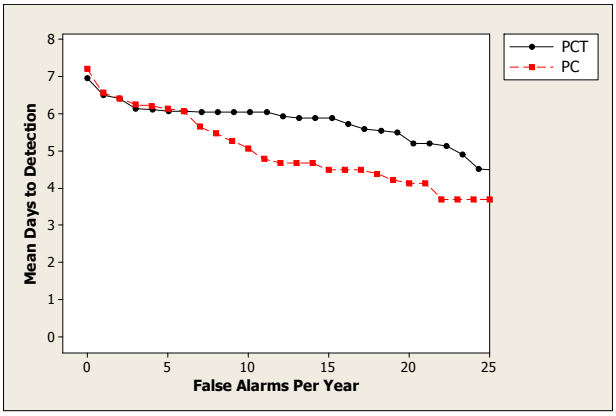
(a) Flu (linear)

(b) Flu (quadratic)

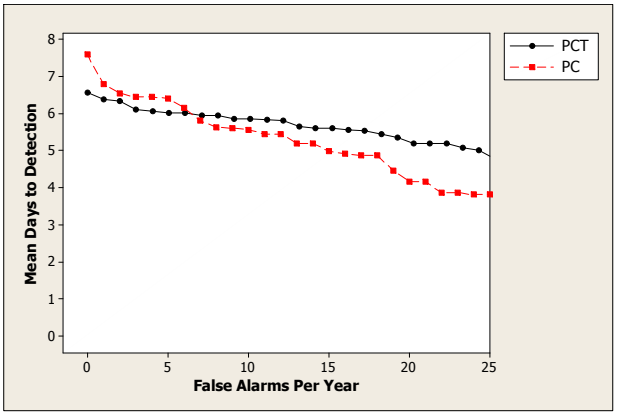
(c) Flu (cubic)

(d) Flu (step)

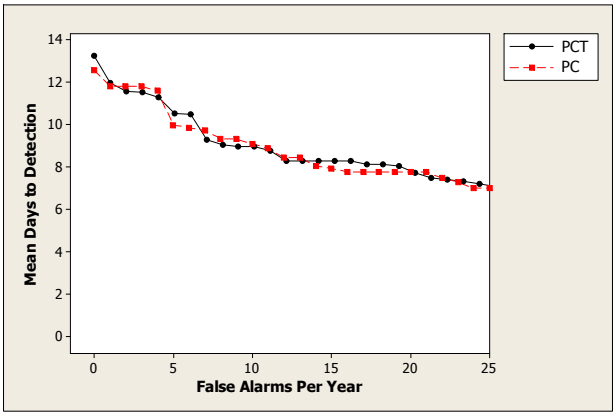
Figure 5.18: The relative performances of PCT and PC for linear-increasing, quadratic-increasing, cubic-increasing, and step-function influenza outbreaks.



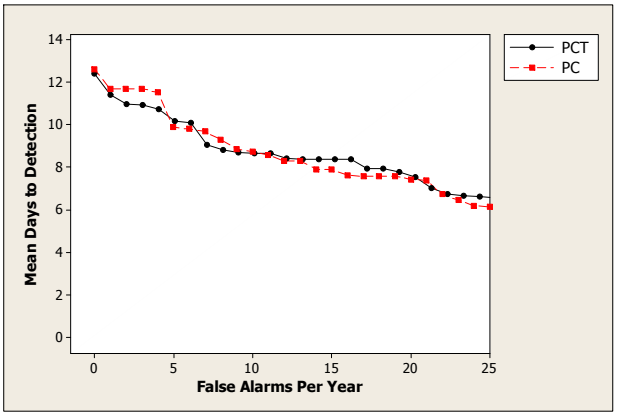
(a) *Cryptosporidium* (25 %)



(b) *Cryptosporidium* (50 %)



(c) Flu (25 %)



(c) Flu (50 %)

Figure 5.19: The relative performances of PCT and PC for different fluctuation values.

substantiate my conjecture that PCT's comparative performance to PC should improve as the daily fluctuation decreases.

5.7.2 Logistic Regression

I performed logistic regression, using the 240 simulated outbreaks as data points, to discover the impact of disease type, fluctuation percentage, subregion size, and epidemic curve function on the relative performances of PCT and PT. The response variable P was given value zero if PC detected the outbreak at least as early as PCT when the false alarm rate was zero, and it was given value one if PCT detected the outbreak earlier than PC when the false alarm rate was zero. The following table shows the variables and their values:

Variable	Value	When the Variable Takes This Value
D	1	Disease is influenza
	2	Disease is *Cryptosporidium*
F	1	Fluctuation is 50%
	2	Fluctuation is 25%
S	1	Subregion shape is 1 by 2 or 2 by 1 (smallest size)
	2	Subregion shape is 2 by 2
	3	Subregion shape is 3 by 2 or 2 by 3 (largest size)
I	1	Linear increasing function
	2	Quadratic increasing function
	3	Cubic increasing function
P	0	PC detects outbreak at least as early as PCT at $FAR = 0$
	1	PCT detects outbreak earlier than PC at $FAR = 0$

The following table shows the results of logistic regression:

Variable	Coefficient	p-value
Constant	-4.87292	0.001
D	2.79111	0.000
F	-0.870383	0.042
S	-0.457231	0.071
I	0.213026	0.395

The results for disease (D) and fluctuation (F) are consistent with the results in Section 5.7.1. Based on the AMOC curves in that section, PCT appears to perform relatively better (at a false positive rate of 0) in the case of *Cryptosporidium* outbreaks. In our regression analysis the coefficient for D is positive, indicating performance should be better for *Cryptosporidium* outbreaks, and the p-value for that coefficient was reported as 0.000, which means that it is less than 0.0005. This is very significant. Furthermore, based on the AMOC curves in Figure 5.19, PCT appears to perform relatively better when the fluctuation is 50%. In the regression analysis the coefficient for F is negative, indicating performance should be better for a fluctuation percentage of 50%, and the p-value for that coefficient is 0.042, which is moderately statistically significant. The result for epidemic curve function are not the same as those discussed in Section 5.7.1.1. We noted in that section that the AMOC curves gave mild support for the conjecture that PCT should perform the best when the epidemic curve function is linear and the worst when it is cubic. However, in our regression analysis the coefficient for I (epidemic curve function) is positive, indicating the performance of PCT is best for cubic functions. However, the result is not significant, having a p-value of 0.395. We noted in Section 5.7.1.1 that the AMOC curves results were not compelling either. Based on both results, we cannot conclude any relationship between epidemic curve function and the performance of PCT. Finally, in the regression analysis the coefficient for S (subregion size) was negative, indicating that PCT should perform better for smaller subregions. The p-value for this result was 0.071, which is close to being considered significant. This result seems reasonable because we have fewer outbreak cases in smaller subregions, which means any one day of data might not contain sufficient outbreak cases to allow detection. PC only looks at one day of data, while PCT looks at several days of data.

Example 5.3. *Let us investigate the best-case scenario for PCT according to our logistic regression results. Suppose we have a Cryptosporidium outbreak ($D = 2$), a fluctuation percentage of 50% ($F = 1$), a subregion of size 1 by 2 ($S = 1$), and a cubic epidemic curve function ($I = 3$). Then according to our regression analysis,*

$$\begin{aligned} logit &= -4.87292 + 2.79111D - 0.870383F - 0.457231S + 0.213026I \\ &= -4.87292 + 2.79111 \times 2 - 0.870383 \times 1 - 0.457231 \times 1 + 0.213026 \times 3 \\ &= 0.020764, \end{aligned}$$

and therefore our estimate according to the logistic regression model is

$$\begin{aligned} P &= P(PCT \text{ detects the outbreak earlier at } FAR = 0) \\ &= \frac{e^{-0.020764}}{1 + e^{-0.020764}} = 0.49481. \end{aligned}$$

Note that this is the probability that PCT will perform better than PC. The remaining probability is allocated to PCT performing the same as PC and to PCT performing worse that PC.

Example 5.4. *Next we investigate the worst-case scenario for PCT. Suppose we have an influenza outbreak ($D = 1$), a fluctuation percentage of 25% ($F = 2$), a subregion of size 3 by 2 ($S = 3$), and a linear epidemic curve function ($I = 1$). Then*

$$\begin{aligned} logit &= -4.87292 + 2.79111D - 0.870383F - 0.457231S + 0.213026I \\ &= -4.87292 + 2.79111 \times 1 - 0.870383 \times 2 - 0.457231 \times 3 + 0.213026 \times 1 \\ &= -4.\,981\,24, \end{aligned}$$

and therefore our estimate according to the logistic regression model is

$$\begin{aligned} P &= P(PCT \text{ detects the outbreak earlier at } FAR = 0) \\ &= \frac{e^{-4.\,981\,24}}{1 + e^{-4.\,981\,24}} = 0.0068187. \end{aligned}$$

6.0 CONCLUSIONS AND FUTURE RESEARCH

6.1 CONCLUSIONS

This dissertation introduced a high-level Bayesian network architecture representing a class of Bayesian network models for spatial event surveillance called BayesNet-S. It further introduced a high-level Bayesian network architecture representing a class of Bayesian network models for temporal event surveillance called BayesNet-T. These architectures were combined into one high-level Bayesian network architecture representing a class of Bayesian network models for spatio-temporal event surveillance called BayesNet-ST.

As discussed in Section 2.3.1.2, PC is an entity-based, non-spatial, non-temporal outbreak detection system that uses a Bayesian network model. This dissertation extended PC to be a BayesNet-S model that is an entity-based, non-temporal, *spatial* outbreak detection system, which is called PCS. PC was also extended to be a BayesNet-T model that is an entity-based, *temporal*, non-spatial outbreak detection system called PCT. These two models were combined resulting in an entity-based, *temporal*, *spatial* outbreak detection system called PCTS. The lattice in Figure 6.1 shows a hierarchical structure for these systems. In that lattice, the system at each node is an enhancement of the system at each of the node's children.

I hypothesized that the system at each node in the lattice in Figure 6.1 in some way improves event surveillance relative to the system at the node's children. The four specific hypotheses were as follows:

1. PCS is an improvement over PC in that, on the average, it is able to detect earlier that an outbreak is occurring. Second, PCS can accurately locate the subregion in which an

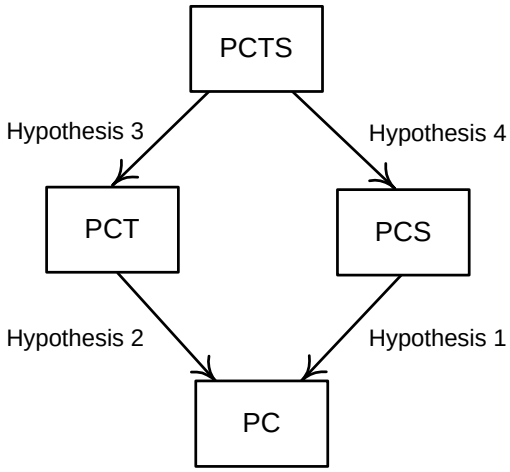


Figure 6.1: A hierarchy of systems and hypotheses.

outbreak is occurring when the outbreak is restricted to a subregion of the monitored region.

2. PCT is more stable than PC in that once an outbreak is detected, PCT is better at maintaining the detection signal on future days. Furthermore, PCT can accomplish this without adversely affecting initial detection capability.
3. PCTS is an improvement to PCT in that, on the average, it can detect earlier that an outbreak is occurring.
4. PCTS is an improvement to PCS in that once an outbreak is detected, PCTS can better maintain the detection signal on future days. Furthermore, PCTS can accomplish this without adversely affecting initial detection capability.

Section 4.3.2 provided results that served to validate Hypothesis 1. In simulated outbreaks using semi-synthetic data, which occurred in spatial subregions, PCS detected the outbreaks much earlier than PC. Furthermore, the results indicated that PCS can detect the correct subregion early in the outbreak. Section 4.3.3 discussed results validating Hypothesis 2. In simulated outbreaks using semi-synthetic data, PCT maintained the detection signal much better than PC. Furthermore, this was accomplished without sacrificing initial

detection capability. In Section 4.3.4 results substantiating Hypothesis 3 were presented. In simulated outbreaks using semi-synthetic data, which occurred in spatial subregions, PCTS detected the outbreaks much earlier than PCT. Section 4.3.5 provided results substantiating Hypothesis 4. In simulated outbreaks using semi-synthetic data, PCTS maintained the detection signal much better than PCS in the case of both *Cryptosporidium* and influenza outbreaks. Furthermore, PCTS performed as well or better than PCS at initial detection of the *Cryptosporidium* outbreaks, while PCTS performed slightly worse that PCS at initial detection of the influenza outbreaks.

The results in Section 5.5.2 indicate that, for the types of outbreaks simulated, PCTS may perform moderately better than SaTScan^{TM}-MT at outbreak detection. A result that was puzzling was that PCTS detected *Cryptosporidium* outbreaks much better than SaTScan^{TM}-MT, but detected influenza outbreaks about the same. The results in Section 5.5.2 also indicate that, for the types of outbreaks simulated, PCTS appears to be better at subregion detection than SaTScan^{TM}-MT.

The results in Section 5.3.2 indicated the PCT is better at outbreak detection than the classic outbreak detection algorithm CUSUM.

At the beginning of Section 1.1.6, I conjectured that we may obtain better results if we modeled the relationships among the event of interest, and the observable events using a Bayesian network instead of using a summary statistic. The results concerning the comparisons of PCTS and PCT to SaTScan^{TM}-MT and CUSUM respectively lend support to this conjecture.

As mentioned in Section 1.2.4, applications of the architectures developed in this thesis extend beyond biosurveillance. My spatial architecture is applicable to many types of anomaly detection including medical imaging for the purpose of pathology detection. My temporal architecture is applicable to any type of monitoring that concerns a system which changes over time. For example, it may be used in a medical expert system, which is deployed in an intensive care unit (ICU), and which monitors changes in a patient's condition over time.

6.2 FUTURE RESEARCH

Future research could provide further comparisons of PCTS, SaTScan^{TM}-MT, and the multivariate version of BSS. Specifically, we could investigate the following:

1. We saw that PCTS performed much better than SaTScan^{TM}-MT at detecting *Cryptosporidium* outbreaks, but not at detecting influenza outbreaks. We could compare detection capability for all 12 types of outbreaks to better evaluate the relative strengths of the systems.
2. An advantage of PCTS is that it is designed to detect 12 different types of outbreaks, while SaTScan^{TM}-MT can only focus on one particular type of outbreak. We would need to run 12 versions of SaTScan^{TM}-MT to realize the generality of PCTS. However SaTScan^{TM}-MT and BSS could consider all 54 chief complaints, and report the occurrence of a non-specific outbreak. We could run a set of simulations that included outbreaks of all 12 types, and compare the abilities of the systems to detect a non-specific outbreak.
3. We could investigate how well the systems detect an outbreak such as Salmonella, which is not one of the outbreaks considered by PCTS. In these comparisons we would use the general-purpose versions of SaTScan^{TM}-MT and BSS just mentioned.
4. We could simulate more than one outbreak occurring concurrently. We could compare how well each system determined a non-specific outbreak. In the case of PCTS we could investigate whether the simulated outbreaks are the ones it considers most probable.

The research just discussed all concerns additional experiments. Future research on modeling could investigate developing a different BayesNet-ST extension of PC. For example, we could investigate the development of a model that does not assume a patient's ED visits on different days are probabilistically independent. Future research on modeling could also investigate extending a different Bayesian network than the one in PC to a BayesNet-ST model. One possibility would be to extend a version of PC that included additional properties of various outbreaks. For example, if there is a *Cryptosporidium* outbreak, there is likely to be contaminated water, and if there is an anthrax outbreak it is likely that spores will be

discovered. Another possibility would be extend a version of PC that does not assume that disease outbreak types are mutually exclusive.

BIBLIOGRAPHY

Anderson, D., Sweeny. D., and Williams, T., 2005, *Statistics for Business and Economics*, South-Western, Mason, OH.

Assuncão, R., M. Costa, Tavares, A., and Ferreira, S., 2006, "Fast Detection of Arbitrarily Shaped Disease Clusters," *Statistics in Medicine*, Vol. 25.

Berry, D.A., 1996, *Statistics: A Bayesian Perspective*, Wadsworth, Belmont, CA.

Bos, T., and Fetherston, T.A., 1992, "Market Model Nonstationarity in the Korean Stock Market," in Rhee, S.G., and Chang, R.P. (Eds.): *Pacific-Basin Capital Markets Research*, Vol. 3, Elsevier, North-Holland, Amsterdam.

Box, G., Jenkins, G., and Reinsel, G., 1994, *Time Series Analysis: Forecasting and Control*, Prentice Hall, Englewood Cliffs, NJ.

Bravata, D.M., McDonald, K.M., Smith, W.M., Rydzak, C., Szeto, H., Buckeridge, D.L., Haberland, C., and Owens, D.K., 2004, "Systematic Review: Surveillance Systems for Early Detection of Bioterrorism-Related Diseases," *Annals of Internal Medicine,* Vol. 140, No. 11.

Buckeridge, D.L., 2007, "Outbreak Detection Through Automated Surveillance: A Review of the Determinants of Detection," *Journal of Biomedical Informatics*, Vol. 40, No. 4.

Buckeridge D.L., Burkom, H., Campbell, M., Hogan, W.R., and Moore, A.W., 2005a, "Algorithms for Rapid Outbreak Detection: a Research Synthesis," *Journal of Biomedical Informatics*, Vol. 38, No. 2.

Buckeridge, D.L., Switzer, P., Owens, D., Siegrist, D., Pavlin, J., and Musen. M., 2005b, "An Evaluation Model for Syndromic Surveillance: Assessing the Performance of a Temporal Algorithm," *MMWR Morb. Mortal. Wkly. Rep. 2005 Aug 26.*

Buntine, W., 1994, "Operations for Learning with Graphical Models," *Journal of Artificial Intelligence Research*, Vol. 2.

Burges, C., 1998, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, Vol. 2, No. 2.

Burkom, H.S., Elbert, Y., Feldman, A., and Lin, J., 2004, "Role of Data Aggregation in Biosurveillance Detection Strategies with Applications from ESSENCE," *MMWR Morb. Mortal. Wkly. Rep. 2004 Sep 24.*

Burkom, H.S., Murphy, S., Coberly, J, and Hurtmullen, K., 2005, "Public Health Monitoring Tools for Multiple Data Streams," *MMWR Morb. Mortal. Wkly. Rep. 2005 Aug 26.*

Burkom, H.S., Murphy. S., and Shmueli, G., 2007, "Automated Time Series Forecasting for Biosurveillance," *Statistics in Medicine*, Vol., 26, No. 22.

Castillo, E., Gutiérrez, J.M., and Hadi, A.S., 2007, *Expert Systems and Probabilistic Network Models*, Springer-Verlag, New York.

Chapman, W.W., Christensen, L.M., Wagner, M.M., Haug, P.J., Ivanov, O., Dowling, J.N., and Olszewski, R.T., 2005, "Classifying Free-Text Triage Chief Complaints into Syndromic Categories with Natural Languages Processing," *Artificial Intelligence in Medicine*, Vol. 33, No. 1.

Chapman, W.W., Dowling, J.N., and Wagner, M.M., 2004, "Fever Detection from Free-Text Clinical Records for Biosurveillance," *Journal of Biomedical Informatics*," Vol. 37, No. 2.

Cooper, G.F., Dash, D.H., Levander, J.D., Wong, W.K., Hogan, W.R., and Wagner, M.M., 2004, "Bayesian Biosurveillance of Disease Outbreaks," In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence.* AUAI Press, Arlington, VA.

Cooper, G.F., Dowling, J.N., Lavender, J.D., and Sutovsky, P., 2007, "A Bayesian Algorithm for Detecting CDC Category A Outbreak Diseases from Emergency Department Chief Complaints," *Advances in Disease Surveillance*, Vol. 2, No. 45.

DeGroot, M.H., 1970, *Optimal Statistical Decisions*, McGraw-Hill, New York.

Duczmal, L., and Assuncão, R., 2004, "A Simulated Annealing Strategy for the Detection of Arbitrarily Shaped Spatial Clusters," *Computational Statistics & Data Analysis*, Vol. 45, No. 2.

Duczmal, L., Cançado, André, Takahashi, R., 2008, "Delineation of Irregularly Shaped Disease Clusters Through Multiobjective Optimization," *Journal of Computational & Graphical Statistics*, Vol. 17, No. 1.

Dwass, M., 1957, "Modified Randomization Tests for Nonparametric Hypotheses," *Annals of Mathematical Statistics*, Vol. 28.

Espino, J.U., Dowling, J., Levander, J., Sutovsky, P., Wagner, M.M., and Cooper, G.F., 2007, "Syco: a Probabilistic Machine Learning Method for Classifying Chief Complaints into Symptom and Syndrome Categories," *Advances in Disease Surveillance 2007*, Vol. 2, No. 5.

Fawcett, T., and Provost, F., 1999, "Activity Monitoring: Noticing Interesting Changes in Behavior," In *Proceedings of the Fifth SIGKDD Conference on Knowledge Discovery and Data Mining*, ACM Press, San Diego, CA.

Fienberg, S.E., and Shmueli, G., 2005, "Statistical Issues and Challenges Associated with Rapid Detection of Bio-terrorist Attacks,"

Statistics in Medicine, Vol. 24, No. 4.

Hamilton, J., 1994, *Time Series Analysis*, Princeton University Press, Princeton, NJ.

Heckerman, D., 1996, "A Tutorial on Learning with Bayesian Networks," Technical Report # MSR-TR-95-06, Microsoft Research, Redmond, WA.

Hjalmars, U., Kulldorff, M., Gustafsson, G., and Nagarwalla, N., 1996, "Childhood Leukemia in Sweden Using GIS and a Spatial Scan Statistic for Cluster Detection," *Statistics in Medicine*, Vol. 15.

Henrion, M., Pradhan, M., Del Favero, B., Huang, K., Provan, G., and O'Rorke, P., 1996, "Why is Diagnosis Using Belief Networks Insensitive to Imprecision in Probabilities?," in Horvitz, E., and Jensen, F. (Eds.): *Uncertainty in Artificial Intelligence; Proceedings of the Twelfth Conference*, Morgan Kaufmann, Burlington, MA.

Hogan, W., Cooper, G.F., Wallstrom, G., and Wagner, M., 2007, "The Bayesian Aerosol Release Detector: an Algorithm for detecting and Characterizing outbreaks Caused by an Atmospheric Release of Bacillus Anthracis," *Statistics in Medicine,* Vol. 26.

Hutwagner, L., Thompson, W., Seeman, G.M., and Treadwell, T., 2003, "The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS)," *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, Vol. 80 (Supplement 1).

Jensen, F.V., 1997, *An Introduction to Bayesian Networks*, Springer-Verlag, New York.

Jensen, F.V., and T.D. Neilsen, 2007, *Bayesian Networks and Decision Graphs*, Springer-Verlag, New York.

Jiang, X., 2006, "A Bayesian Network for Predicting an Epicurve," In *Proceedings of Syndromics 2006*, Baltimore, Maryland.

Jiang, X.,and Wallstrom, G.L., 2006, "A Bayesian Network for Outbreak Detection and Prediction," In *Proceedings of AAAI-06*, Boston, MA.

Jiang, X., and G.F. Cooper, 2007, "A Recursive Algorithm for Spatial Cluster Detection," *Proceedings of AMIA 2007 Annual Symposium*, Chicago, IL, November, 2007.

Jung, I., M. Kulldorff, and Klassen, A., 2007, "A Spatial Scan Statistic for Ordinal Data," *Statistics in Medicine*, Vol. 26.

Kaufmann, A., Meltzer, M., and Schmid, G., 1997, "The Economic Impact of a Bioterrorist Attack: Are Prevention and Postattack Intervention Programs Justifiable," *Emerging Infectious Diseases*, Vol. 3.

Kjaerulff, U.B., and Madsen, A.L., 2008, *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*, Springer-Verlag, New York.

Kullback, S., and R.A. Leibler, 1951, "On Information and Sufficiency," *Annals of Mathematical Statistics*, Vol. 22.

Kulldorff, M., 1997, "A Spatial Scan Statistic," *Communications in Statistics: Theory and Methods*, Vol. 26, No. 6.

Kulldorff, M., 1999, "Spatial Scan Statistics: Models, Calculations, and Applications," in Glaz and Balakrishnan (Eds.): *Scan Statistics and Applications*, Birkhauser, Boston, MA.

Kulldorff, M., 2001, "Prospective Time Periodic Geographical Disease Surveillance Using a Scan Statistic," *J. R. Statist. Soc. A*, Vol. 164.

Kulldorff, M., 2004, "Satscan v. 4.0: Software for the Spatial and Space-time Scan Statistics," Technical Report, Information Management Services, Inc.

Kulldorff. M., Fang, Z., and Walsh, S.J., 2003, "A Tree-Based Scan Statistic for Database Disease Surveillance," *Biometrics*, Vol. 59.

Kulldorff, M., Feuer, E.J., Miller, B.A., and Freedman, L.S., 1997, "Breast Cancer Clusters in the Northeast United States: a Geographical Analysis," *American Journal of Epidemiology*, Vol. 146, No. 2.

Kulldorff, M., Heffernan, R., Hartman, J., Assunco, R., and Mostashari, F., 2005, " Space-Time Permutation Scan Statistic for Disease Outbreak Detection," *PLoS Medicine*, Vol. 2, No. 3.

Kulldorff, M., Mostashari, F., Luiz, D., Yih, K., Kleinman, K., and Platt, R., 2007, "Multivariate Scan Statistics for Disease Surveillance," *Statistics in Medicine*, Vol. 26.

Kulldorff, M., and Nagarwalla, N., 1995, "Spatial Disease Clusters: Detection and Inference," *Statistics in Medicine*, Vol. 14.

Kurki, I., and Saarinen, J., 2006, "Detection of Irregular Spatial Structures," *Spatial Vision*, Vol. 19, No. 5.

Last, J.M. 2000, *A Dictionary of Epidemiology*, Oxford University Press, New York.

Le Strat, Y., and Carrat, F., 1999, "Monitoring Epidemiological Surveillance Data using Hidden Markov Models," *Statistics in Medicine*, Vol. 18.

Lombardo, J., Burkom, H., Elbert, E., Magruder, S., Lewis, S.H., Loschen, W., Sari, J., Sniegosk,i C., Wojcik, R., and Pavlin, J., 2003, "A Systems Overview of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II)," *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, Vol. 80 (Supplement 1).

Lu, H, Zeng, D., Trujillo, L., Komatsu, K., and Chen, H., 2008, "Ontology-Enhanced Automatic Chief Complaint Classification for Syndromic Surveillance," *Journal of Biomedical Informatics*, Vol. 42, No. 2.

Mikosz, C.A., Silva, J., Black, S., Gibbs, G., and Cardenas. I, 2004 "Comparison of Two Major Emergency Department-Based Free-Text Chief-Complaint Coding Systems," *MMWR Morb. Mortal. Wkly. Rep. 2004 Sep 24.*

Moore, A., 2001a, "A Powerpoint Tutorial on Bayes Nets," available at http://www.cs.cmu.edu/~awm/781/timetable.html.

Moore, A., 2001b, "A Powerpoint Tutorial on Hidden Markov Models," available at http://www.cs.cmu.edu/~awm/781/timetable.html.

Moore, A., 2001c, "A Powerpoint Tutorial on Support Vector Machines," available at http://www.cs.cmu.edu/~awm/781/timetable.html.

Moore, A., Anderson, B., Das, K., and Wong, W.K., 2006, "Combining Multiple Signals for Biosurveillance," In Wagner, M. (Ed.): *Handbook of Biosurveillance*, Elsevier, New York.

Moore, A.W., Cooper, G.F., Tsui, F-C, and Wagner, M.M., 2003, "Summary of Biosurveillance-Relevant Statistical and Data Mining Technologies," RODS Technical Report.

Mostashari, F., Kulldorff, M., Hartman, J.J., Miller, J.R., Kulasekera, V., 2003, "Dead Bird Clustering: A Potential Early Warning System for West Nile Virus Activity," *Emerging Infectious Diseases*, Vol. 9.

Naus, J.J., 1965, "The Distribution of the Size of the Maximum Cluster of Points on the Line," *Journal of the American Statistical Association*," Vol. 60.

Neapolitan, R.E., 1990, *Probabilistic Reasoning in Expert Systems*, Wiley, New York.

Neapolitan, R.E., 2004, *Learning Bayesian Networks*, Prentice Hall, Upper Saddle River, NJ.

Neapolitan, R.E., 2008, "A Polemic for Bayeisan Statistics," in Holmes, D. and Jain, L. (Eds.): *Innovations in Bayesian Networks*, Springer-Verlag, Berlin Heidelberg.

Neill, D.B., and Cooper, G..F., 2008, "A Multivariate Bayesian Scan Statistic for Early Event Detection and Characterization," Technical Report, School of Computer Science, Carnegie Mellon University.

Neill, D.B., and Moore, A.W, 2004, "Rapid Detection of Significant Spatial Clusters," In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Neill, D.B., Moore, A.W., and Cooper, G.F., 2005a, "A Bayesian Spatial Scan Statistic," *Advances in Neural Information Processing Systems (NIPS)* ,Vol. 18.

Neill, D.B., Moore, A.W., and Cooper, G.F., 2007, "A Multivariate Bayesian Scan Statistic," *Advances in Disease Surveillance*, Vol. 2, No. 60.

Neill, D.B., Moore, A.W., Sabnani, M., and Daniel, K., 2005b, "Detection of Emerging Space-Time Clusters," *Proceedings of 11th ACM SIGKDD International Conference on Knowledge Discovery and Mining*, Chicago, Illinois.

Olszewski, R.T., 2003, "Bayesian Classification of Triage Diagnoses for the Early Detection of Epidemics," *Proceedings of the 16th International FLAIRS Conference*, Menlo Park, California.

Patil, G.P., and Taillie, C., "Upper Level Set Scan Statistic for Detecting Arbitrarily Shaped Hotspots," *Environmental and Ecological Statistics*, Vol. 11.

Pearl, J., 1988, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo, California.

Rabiner, L.R., 1989, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of IEEE*, Vol. 77, No. 2.

Reis, B.R., and Mandl, K.D., 2003, "Time Series Modeling for Syndromic Surveillance," *BMC Medical Informatics and Decision Making*, Vol. 3, No. 2.

Reis, B.Y., Pagano, M., and Mandl, K.D., 2003, "Using Temporal Context to Improve Biosurveillance," *PNAS*, Vol. 100, No. 4.

Reis, B.Y., Kohane, I.S., and Mandl, K,D., 2007, "An Epidemiological Network Model for Disease Outbreak Detection," *PLoS Medicine*, Vol. 4, No. 6.

Serfling, R.E, 1963, "Methods for Current Statistical Analysis of Pneumonia-Influenza Deaths," *Public Health Reports*, Vol. 28.

Shmueli, G., and Fienberg, S., 2006, "Current and Potential Statistical Methods for Monitoring Multiple Data Streams for Biosurveillance," in Wilson, A., Wilson, G.D., and Olwell, D. (Eds.): *Statistical Methods in Counterterrorism*, Springer, New York.

Soneson, C., and Bock, D., 2003, "A Review and Discussion of Prospective Statistical Surveillance in Public Health," *J.R. Stat. Soc. A*, Vol. 166.

Spiegelhalter, D., 1998, "Bayesian Graphical Modelling: A Case-Study in Monitoring Health Outcomes," *Applied Statistics*, Vol. 47.

Stirling R., Aramini, J., Ellis, A., Gillien, L., Meyers, R., Flevry, M., and Werker, D., 2001, "Waterborne Cryptosporidiosis Outbreak, North Battleford, Saskatchewan, Spring 2001," *Health Canada 2001*, Vol. 2.

Sun, L., and Shenoy, P., 2006, "Using Bayesian Networks for Bankruptcy Prediction: Some Methodological Issues," School of Business Working Paper No. 302, University of Kansas, Lawrence, Kansas.

Takahashi, K., Kulldorff, M., Tango, T., and Yih, K., 2008, "A Flexibly Shaped Space-Time Scan Statistic for Disease Outbreak Detection and Monitoring," *International Journal of Health Geographics*, Vol. 7, No. 14.

Tsui, F.C.R, Wagner, M., Dato, V., and Chang, H.C., 2001, "Value of ICD-9-Coded Chief Complaints for Detection of Epidemics," in *Symposium of the Journal of American Medical Informatics Association.*

Tsui, F.C.R, Espino, J.U., Dato, V.M., Gesteland, P.H., Hutman, J., and Wagner, M.M., 2003, "Technical description of RODS: a Real-Time Public Health Surveillance System." *Journal of the American Medical Informatics Association*, Vol. 10, No.5.

Turnball, B.W., Iwano, E.J., Burnett, W.S., Howe, H.L., and Clark, L.C., 1990, "Monitoring for Clusters of Disease: Application to Leukemia Incidence in Upstate New York," *American Journal of Epidemiology*, Vol. 132.

von Mises, R., 1919, "Grundlagen der Wahrscheinlichkeitsrechnung," *Mathematische Zeitschrift*, Vol. 5.

Wagner, M.M., Tsui, F.C., Espino, J.U., Dato, V.M., Sittig, D.F., Caruana, R.A., McGinnis, L.F., Deerfield, D.W., Druzdzel, M.J., and Fridsma, D.B., 2001, "The Emerging Science of Very Early Detection of Disease Outbreaks," *Journal of Public Health Management and Practice*, Vol. 7.

Wieland, S.C., Brownstein, J.S., Berger B., and Mandl, K.D., 2007, "Density-Equalizing Euclidean Minimum Spanning Trees for the Detection of all Disease Cluster Shapes," *Proceedings of the National Academy of Sciences*, May 22, 2007.

Wong, W.K., and Moore, A., 2006, "Classical Time Series Methods for Biosurveillance," In Wagner, M. (Ed.): *Handbook of Biosurveillance*, Elsevier, New York, NY.

Appendix A

Under certain assumptions the frequentist's p-value for the null hypothesis H_0 is equal to the Bayesian's posterior probability of H_0. I show this first for the case where the mean is unknown and the variance is known; then I address the case where both the mean and the variance are unknown.

Unknown Mean and Known Variance

Suppose X is normally distributed with unknown mean and known precision r (the precision is one divided by the variance). Suppose further that we represent our belief concerning the unknown mean with a random variable A, which is normally distributed with mean μ and precision v. The probability distribution of X is a relative frequency distribution, while the probability distribution of A is our subjective probability concerning the value of X's mean. Then the prior density function of A is

$$\rho_A(a) = \text{NormalDen}(a; \mu, 1/v).$$

Let our $Data$ consists of n values $x_1, x_2, \ldots x_n$ of the random variable X, and set

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

(Neapolitan, 2004, p. 394) shows that the posterior density function of A is

$$\rho_A(a|Data) = \text{NormalDen}(a; \mu^*, 1/v^*),$$

where

$$\mu^* = \frac{v\mu + nr\bar{x}}{v + nr} \qquad \text{and} \qquad v^* = v + nr. \tag{6.1}$$

If we model prior ignorance concerning the value of the unknown mean by assuming that the prior density function of A is the improper uniform density function over the whole real line, (Neapolitan, 2004, p. 396) shows further that the posterior density function of A

$$\rho_A(a|Data) = \text{NormalDen}(a; \bar{x}, 1/nr). \tag{6.2}$$

Intuitively, this result can be obtained by taking the limit as v approaches 0 of the expressions in Equation 6.1. Writing Equation 6.2 using the variance, we have that

$$\rho_A(a|Data) = \text{NormalDen}(a; \bar{x}, \sigma^2/n).$$

Let us model prior ignorance in this way, and suppose that our null hypothesis is as follows:

$$H_0 : a \leq \xi.$$

Then

$$\begin{aligned} P(H_0|Data) &= \int_{-\infty}^{\xi} \text{NormalDen}(a; \bar{x}, \sigma^2/n)da \\ &= \text{NormalDist}(\xi; \bar{x}, \sigma^2/n), \end{aligned}$$

where NormalDist denotes the Normal cumulative distribution function.

If we use the standard notation of the z score, and set

$$z = \frac{\xi - \bar{x}}{\sigma/\sqrt{n}},$$

then

$$P(H_0|Data) = \text{NormalDist}\left(\frac{\xi - \bar{x}}{\sigma/\sqrt{n}}; 0, 1\right).$$

The p-value for H_0 is

$$\begin{aligned} p &= \int_{\bar{x}}^{\infty} \text{NormalDen}(x; \xi, \sigma^2/n)dx \\ &= 1 - \int_{-\infty}^{\bar{x}} \text{NormalDen}(x; \xi, \sigma^2/n)dx \\ &= 1 - \text{NormalDist}(\bar{x}; \xi, \sigma^2/n). \end{aligned}$$

If we use the standard notation of the z score and set

$$z = \frac{\bar{x} - \xi}{\sigma/\sqrt{n}},$$

then

$$\begin{aligned} p &= 1 - \text{NormalDist}\left(\frac{\bar{x}-\xi}{\sigma/\sqrt{n}}; 0, 1\right) \\ &= \text{NormalDist}\left(\frac{\xi-\bar{x}}{\sigma/\sqrt{n}}; 0, 1\right) \\ &= P(H_0|Data). \end{aligned}$$

The second to the last result is due to the symmetry of the normal density function.

Unknown Mean and Unknown Variance

Suppose X is normally distributed with unknown mean and unknown precision r, and that we represent our belief concerning the mean and precision with the random variables A and R respectively. Suppose further that we represent our prior belief concerning the value of R using the $\text{GammaDen}\,(r; \alpha, \beta)$ density function, and our prior belief concerning the value of A using the $\text{NormalDen}\,(a; \mu, 1/vr)$ conditional density function. If we model prior ignorance concerning the values of R and A by assuming that the prior density function of R is the improper density function $1/r$ and the prior density function of A is the improper uniform density function over the whole real line, (DeGroot, 1970, p. 195) shows that

$$t = \frac{a-\bar{x}}{s/\sqrt{n}}$$

has the t distribution with $n-1$ degrees of freedom, where

$$s = \left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}\right)^{1/2}.$$

Let us model prior ignorance in this way, and suppose that our null hypothesis is as follows:

$$H_0 : a \leq \xi.$$

We then have that

$$\begin{aligned} P(H_0|Data) &= P(a \leq \xi) \\ &= \int_{-\infty}^{\frac{\xi-\bar{x}}{s/\sqrt{n}}} \text{TDen}(t, n-1)dt \\ &= \text{TDist}\left(\frac{\xi-\bar{x}}{s/\sqrt{n}}, n-1\right), \end{aligned}$$

where TDist denotes the t cumulative distribution function.

The frequentist test statistic in this case is given by (see e.g. (Anderson et al., 2005))

$$t = \frac{\bar{x} - \xi}{s/\sqrt{n}}.$$

The p-value for H_0 is then

$$\begin{aligned} p &= \int_{\frac{\bar{x}-\xi}{s/\sqrt{n}}}^{\infty} \mathrm{TDen}\left(t, n-1\right) dt \\ &= 1 - \int_{-\infty}^{\frac{\bar{x}-\xi}{s/\sqrt{n}}} \mathrm{TDen}\left(t, n-1\right) dt \\ &= 1 - \mathrm{TDist}\left(\frac{\bar{x}-\xi}{s/\sqrt{n}}, n-1\right) \\ &= \mathrm{TDist}\left(\frac{\xi-\bar{x}}{s/\sqrt{n}}, n-1\right) \\ &= P(H_0|Data). \end{aligned}$$

The second to the last equality is because of the symmetry of the t density function.