ABSTRACT

Title of Dissertation: ANOMALY DETECTION IN TIME SERIES: THEORETICAL AND PRACTICAL IMPROVEMENTS FOR DISEASE OUTBREAK DETECTION.

Thomas Harvey Lotze,
Doctor of Philosophy, 2009

Dissertation Directed By: Professor Galit Shmueli
Department of Decision, Operations and Information Technologies

The automatic collection and increasing availability of health data provides a new opportunity for techniques to monitor this information. By monitoring pre-diagnostic data sources, such as over-the-counter cough medicine sales or emergency room chief complaints of cough, there exists the potential to detect disease outbreaks earlier than traditional laboratory disease confirmation results. This research is particularly important for a modern, highly-connected society, where the onset of disease outbreak can be swift and deadly, whether caused by a naturally occurring global pandemic such as swine flu or a targeted act of bioterrorism. In this dissertation, we first describe the problem and current state of research in disease outbreak detection, then provide four main additions to the field.

First, we formalize a framework for analyzing health series data and detecting anomalies: using forecasting methods to predict the next day's value, subtracting the forecast to create residuals, and finally using detection algorithms on the residuals. The formalized framework indicates the link between the forecast accuracy of the forecast method and the performance of the detector, and can be used to quantify and analyze the performance of a variety of heuristic methods.

Second, we describe improvements for the forecasting of health data series. The application of weather as a predictor, cross-series covariates, and ensemble forecasting each provide improvements to forecasting health data.

Third, we describe improvements for detection. This includes the use of multivariate statistics for anomaly detection and additional day-of-week preprocessing to aid detection. Most significantly, we also provide a new method, based on the CuScore, for optimizing detection when the impact of the disease outbreak is known. This method can provide an optimal detector for rapid detection, or for probability of detection within a certain timeframe.

Finally, we describe a method for improved comparison of detection methods. We provide tools to evaluate how well a simulated data set captures the characteristics of the authentic series and time-lag heatmaps, a new way of visualizing daily detection rates or displaying the comparison between two methods in a more informative way.

ANOMALY DETECTION IN TIME SERIES: THEORETICAL AND PRACTICAL IMPROVEMENTS FOR DISEASE OUTBREAK DETECTION.

By

Thomas Harvey Lotze

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Professor Galit Shmueli, Chair
Dr. Howard Burkom
Professor Bruce Golden
Professor Wolfgang Jank
Professor Ben Shneiderman
Professor Paul Smith

Foreword

The student was responsible for all relevant aspects of any jointly authored work included in this dissertation.

Dedication

This work is dedicated to my parents, Joan and Michael.

Acknowledgements

We thank Howard Burkom of the Johns Hopkins University's Applied Physics Laboratory, for making the aggregated ED data set, previously authorized by ESSENCE data providers for public use at the 2005 Syndromic Surveillance Conference Workshop, available to us.

This research was performed under an appointment to the U.S. Department of Homeland Security (DHS) Scholarship and Fellowship Program, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and DHS. ORISE is managed by Oak Ridge Associated Universities (ORAU) under DOE contract number DE-AC05- 06OR23100. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of DHS, DOE, or ORAU/ORISE.

The calculations in this dissertation were performed using R, the open source statistical programming language (R Development Core Team, 2009).

Table of Contents

List of Tables

List of Figures

Chapter 1 : Introduction to Biosurveillance

1.1. Biosurveillance

1.1.1. Introduction

In modern biosurveillance, time series of diagnostic and pre-diagnostic health data are monitored for the purpose of detecting disease outbreaks. In general, the data tend to be indirect measures of a disease (as opposed to more traditional diagnostic or clinical data). Examples of pre-diagnostic biosurveillance health data include daily counts of emergency room visits, over-the-counter (OTC) or prescription medication sales, school absences, doctors' office visits, veterinary reports, web searches for disease-related terms, or other data streams that could contain an indication of a disease outbreak. These data are usually collected for a specific region of interest, such as that covered by a public health department. Outbreaks of interest include terrorist-driven attacks, such as a bioterrorist anthrax release, or naturally occurring epidemics, such as an avian or porcine influenza outbreak. In either setting, the goal is to alert public officials and create an opportunity for them to respond in a timely manner. To effectively provide this opportunity, alerts must occur quickly after the outbreak begins, should detect most outbreaks, and have a low false alert rate. There are a host of statistical difficulties in achieving such performance (as described in (Fienberg & Shmueli, 2005, Shmueli & Burkom, 2009)), foremost among them the seasonal, nonstationary, and autocorrelated nature of the health data being monitored. There are also data collection issues such as delayed data transmission or unexpected increases in the number of reporting hospitals. Although current biosurveillance data are

typically monitored at a daily frequency, the methods and results in this dissertation are general and apply to data at other time scales as well.

Our ultimate purpose is to provide early notice of an outbreak based on finding an outbreak signature in the data. We will refer to the outbreak signature as an "outbreak signal" or sometimes simply the "outbreak". However, it should be clear that there is a distinction between the outbreak itself and its manifestation or signature in the monitored data series. For evaluation purposes, algorithms must be evaluated on their ability to detect these outbreak signatures. In this chapter, we first describe the metrics used to evaluate the performance of a biosurveillance algorithm. We then discuss current systems being used in practice, describe the authentic data sets which will be used for algorithm testing throughout this dissertation, and then review the research which has been done on statistical methods for biosurveillance.

1.1.2. A Brief History of Biosurveillance

The purpose of biosurveillance is to understand the health of a population, and in particular to understand the health problems present in the population and how they are progressing through the population. This understanding often leads to investigation of the underlying causes of illness and estimation of the future progression of illness. Thus, biosurveillance is closely related to epidemiology and is sometimes thought of as a sub-field. However, biosurveillance is distinguished by its focus on continual monitoring, using information technology to provide up-to-date quantitative reports, and resulting in timely intervention rather than retrospective analyses.

While epidemiology can trace its origins to Hippocrates' study of the relationships between environmental factors and disease, it only truly developed with the germ theory of disease. John Snow's famous investigation of the 1854 Birmingham cholera epidemic is an early example of epidemiology; by plotting the cholera deaths, he was able to determine the source of the cholera, a contaminated water pump, and intervene (by removing the handle) to stop the outbreak.

Figure 1-1: John Snow's Map of Cholera Deaths
John Snow's map showing deaths from cholera (each marked with a dot) and locations of water pumps (each marked with an X) indicates the link between the Broad Street pump and the cholera epidemic.

Epidemiology is characterized by the use of investigation to determine the link between the root cause of the disease and its appearance in the human population (Green et al., 2000). Epidemiology through the 19th and mid-20th centuries was usually directed at proving the existence or disease-causing role of infectious or environmental agents; a more recent canonical example is the epidemiological studies of lung cancer, such as (Doll & Hill, 1956), leading eventually to the establishment of

tobacco smoke as a contributing factor. As the science of bioinformatics developed and more information on public health became readily available, epidemiology came to use these tools to perform its causal studies.

As these data became more prevalent, it became possible to use them not merely for designed studies, as in an epidemiologic case study, but to regularly record such information and use it for monitoring public health. One can think of biosurveillance as the development of epidemiologic methods for continual health monitoring, rather than post-hoc analyses. Traditional data sources for biosurveillance include laboratory tests, such as those looking for antibodies to specific diseases (such as influenza variants). Such data can be used both for monitoring (biosurveillance) or cause analysis and investigation (epidemiology). Biosurveillance is a natural partner to epidemiology; the ability to find outbreaks is not useful without the ability to track down their cause and determine an appropriate intervention.

Biosurveillance has developed particular prominence in the past ten years mainly due to fears of two scenarios: first, the threat of bioterrorist attacks, where a terrorist group obtains and releases a biological disease agent such as anthrax; and second, the threat of naturally-occurring pandemics with the potential to spread rapidly due to modern transportation and greater human mobility, such as SARS or swine flu. Because of this, the focus shifted to early alerts of disease outbreaks.

As data availability has increased, biosurveillance has become a possible source of situational awareness, with the ability to provide alerts of outbreaks as they happen. This is further enhanced by the inclusion of pre-diagnostic data, data which indicate increases in syndromes for specific diseases or simply more general disease symptoms. Rather than waiting days after the start of infection for laboratory confirmation, pre-diagnostic sources can provide indications of disease which allow public health officials to respond earlier, potentially reducing the impact of the disease and saving lives. While early health indicators such as over-the-counter (OTC) drug sales, emergency department chief complaints, and absentee records do not provide direct indication of disease, but instead simply give an indicator of symptom effect or care-seeking behavior, they are less specific than traditional laboratory reports. However, their ability to give an earlier signal makes their analysis an important tool for public health monitoring. It is in this context that biosurveillance has developed, seeking methods to analyze and report potential disease outbreaks using this challenging but rewarding data source.

1.1.3. Intervention Effects

The principle behind biosurveillance is that by providing early notification of disease outbreaks, public health officials can respond to reduce the severity of the disease impact. However, because we do not know what would have happened without the intervention, it is difficult to measure the effect of any action. Some recent studies attempt to measure that impact on school closures in Hong Kong (Cowling et al., 2008), on influenza immunization (Davis et al., 2008), on measles inoculation (Grais et al., 2007), and on heat wave-related mortality (Josseran et al., 2009). There has

been strong evidence that when intervention is performed in a timely manner, the effect is meaningful.

1.1.4. Performance Evaluation Metrics

Consider a time series of health data, collected periodically. Daily is the most common collection interval, and we use the convention of assuming daily collection throughout the dissertation; however, our theoretical results apply equally well for different intervals. Now consider that we have many such series of the same type; some contain outbreaks, and some do not. What we are looking for in biosurveillance are methods which perform well on many different series. The assumption is that in the future, if the method is used on similar series, it will perform well.

The main metrics used in biosurveillance to evaluate an outbreak detection method are *sensitivity*, *specificity*, and *timeliness*. The first two metrics are widely used in public health. Sensitivity measures how effective a method is at detecting an outbreak, assuming one exists; specificity measures how many false alerts will be generated by that same method; and timeliness measures how quickly, after the start of the outbreak, the method detects. Specificity and sensitivity are closely related to the probability of type I and type II error, respectively; if the probability of type I error is α, then specificity is $1 - \alpha$ and if the probability of type II error is $1 - \beta$, then the sensitivity is β.

In biosurveillance we are considering not simply a single decision on whether the disease outbreak is present, but an alert decision made repeatedly over each day.

Because the decision process is repeated each day, one must consider the specificity as a rate over time during which there is no outbreak. Because outbreaks can last multiple days, an alert can be generated on several potential days and be valid; it is therefore useful to think of the sensitivity as an overall probability of alert during the outbreak. For this reason, to measure these characteristics we use the measures described in (Fricker et al., 2008b), which are closer to those used in statistical process control. For the specific definitions below, consider that we take k series, each with an outbreak, and m series without an outbreak.

Detection Rate: the proportion of outbreaks detected, out of the k series with outbreaks. As k is made arbitrarily large, this measures the per-outbreak probability that there will be an alert sometime during the outbreak. This is also sometimes referred to as True Alert rate (*TA*).

ATFS: the Average Time to False Signal, this is the average number of days until an alert, over the m series without outbreaks. As m is made arbitrarily large, this measures the expected time until a false alert. For implementations which reset after any alert, 1/ATFS will be the average proportion of days with false alerts, given that there is no outbreak. We will sometimes use the term False Alert rate (*FA*) as 1/ATFS.

ATFOS: the Average Time to First Outbreak Signal, this is the expected number of days until an alert is generated, given that the method does eventually alert

during the outbreak signal. We will also sometimes use the term *Delay* for ATFOS, or describe a method's ATFOS performance as its *timeliness*.

The ATFS and Detection Rate are often shown graphically using Receiver Operating Characteristic (ROC) curves. ROC curves plot the Detection Rate on the y-axis for different False Alert levels on the x-axis. Figure 1-2 is an example. An Activity Monitoring Operating Characteristic (AMOC) curve is similar, but measures delay on the y-axis instead of Detection Rate. The area under the ROC curve (AUC) is a common measure of performance, as it sums the algorithms performance over all possible false alert levels. This measure is often restricted to a range of practically useful False Alert levels, in order to compare performance over false alert levels which can be managed by the available resources. Figure 1-3 shows an example over False Alert rates between 1/28 and 1/7.

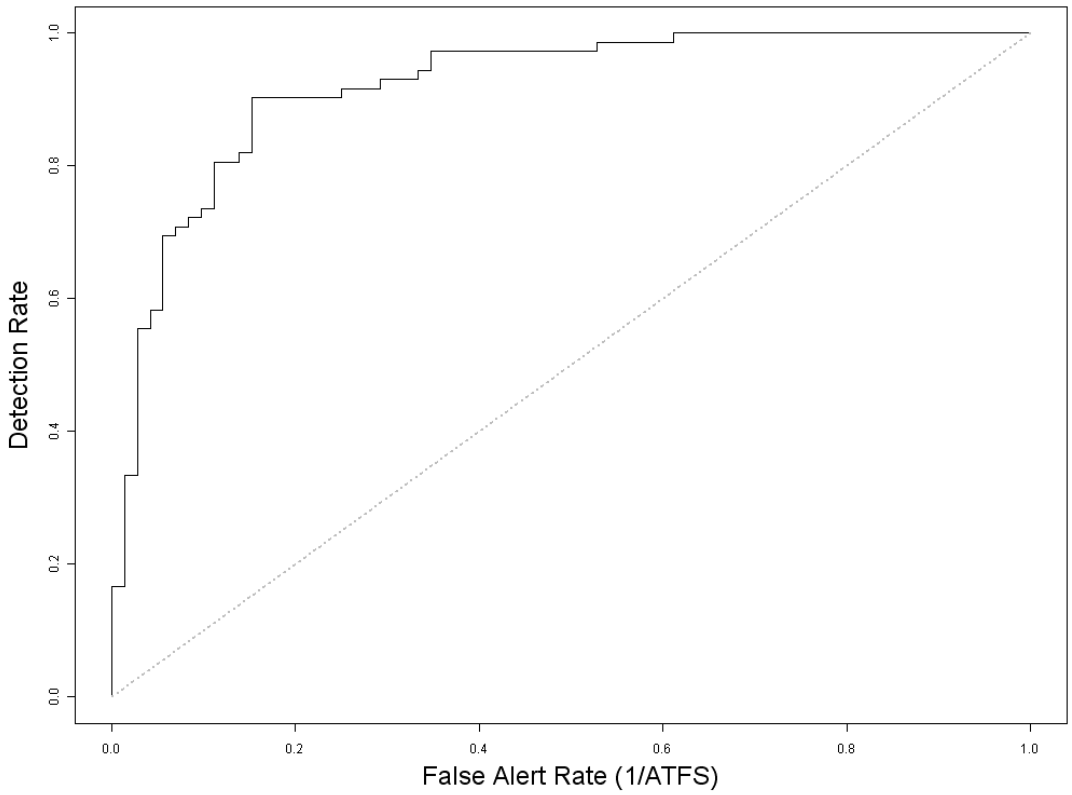

Figure 1-2: ROC Curve Example

A basic ROC curve, showing the Detection Probability of an algorithm, for varying False Alert Rates. Any reasonable algorithm will have a monotonically increasing ROC curve, which reflects the fact that a higher rate of false alerts should allow the algorithm to detect an increased number of actual outbreaks. The diagonal line shows the performance of an algorithm which generates alerts by chance.

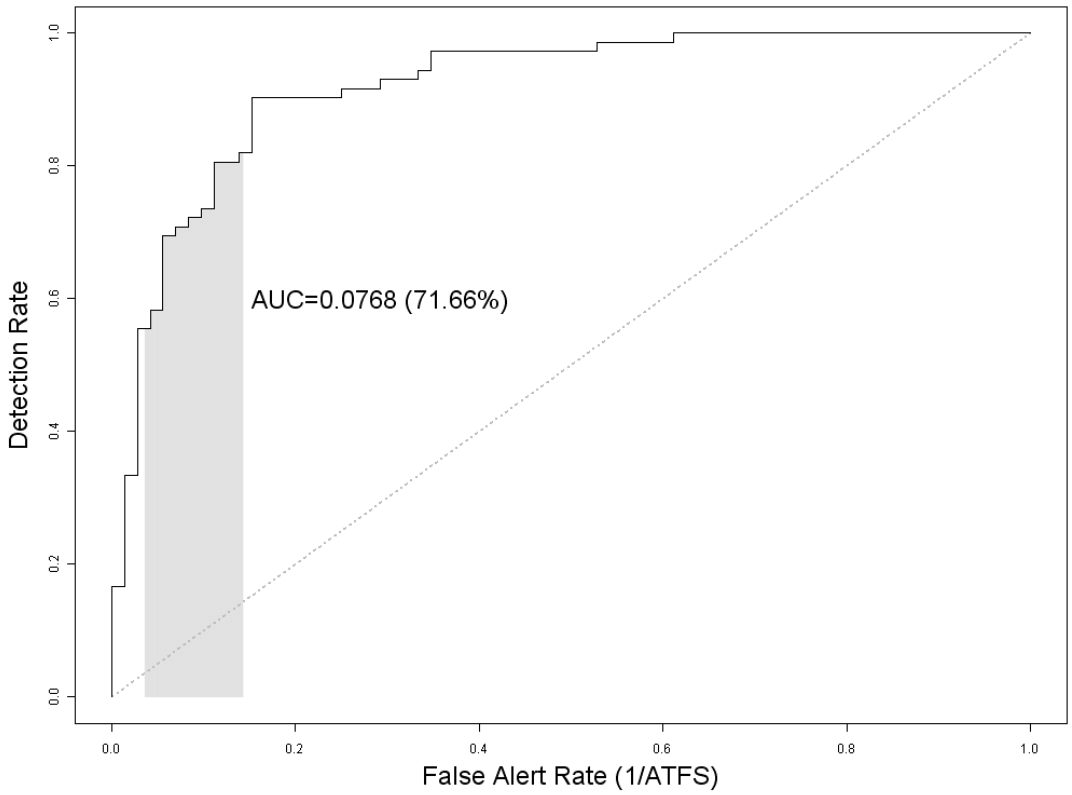

Figure 1-3: AUC Example
An illustration of the AUC for a section of the ROC curve, corresponding to false alert rates of one every 7 days and one every 28 days. An algorithm with a higher AUC will have a higher average Detection Rate over the range of false alert levels.

We note here that the detection performance depends on the outbreak signal itself, as well as on the underlying health data series. In biosurveillance the variety of data sources leads to a variety of baseline behaviors; emergency room respiratory chief complaints may look very different than elementary school absences, even over the same time period and in the absence of a disease outbreak. Furthermore, the exact outbreak signal is unknown. Therefore, it is generally important to consider a variety of baseline time series as well as a variety of outbreak signal shapes and sizes for evaluating algorithm performance. Given the wide array of possibilities, simulation methods, and metrics, it is difficult to make overall claims about the performance of one method versus another. We will discuss how to evaluate simulations in Section

5.2, and discuss the theory for comparing methods using a theoretical framework in Chapter 2.

1.2. Existing Biosurveillance Systems in the United States

We next briefly review the major existing biosurveillance systems in the United States. While other countries have increasingly been developing biosurveillance systems with substantial capabilities and effectiveness, the U.S. systems remain the most prominent. We do not mean to indicate that other systems are not worth consideration, only that focusing on the U.S. systems allows for a salient overview.

1.2.1. RODS

RODS (Real-Time Outbreak and Disease Surveillance) is a program developed by the University of Pittsburgh in 1999 as a monitoring system to detect anthrax outbreaks (Wagner et al., 2003, Tsui et al., 2003). It is now an open source (Espino et al., 2004) general outbreak detection software package, implemented in Java. RODS is now used by hundreds of public health departments, both within the US and internationally. It is still used as a development testbed for further algorithm development by the University of Pittsburgh. Although this research has tapered off in recent years, the open source nature of the project ensures that it will not be lost and can continue to support development.

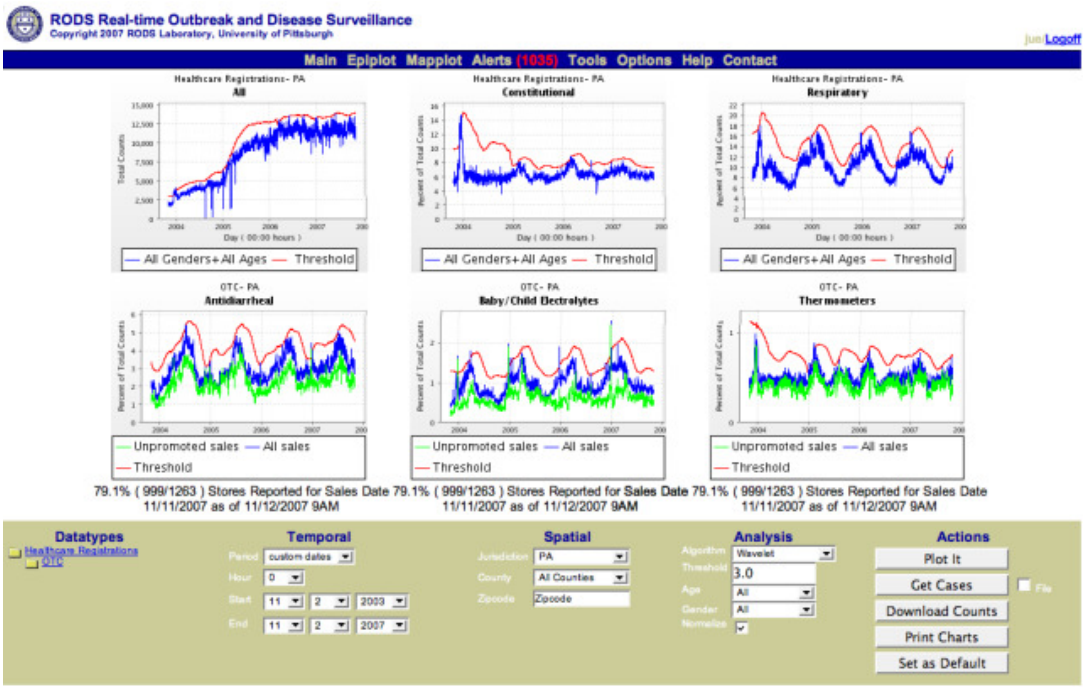

Figure 1-4: RODS Main Visualization
Main visualization screen for the RODS system.

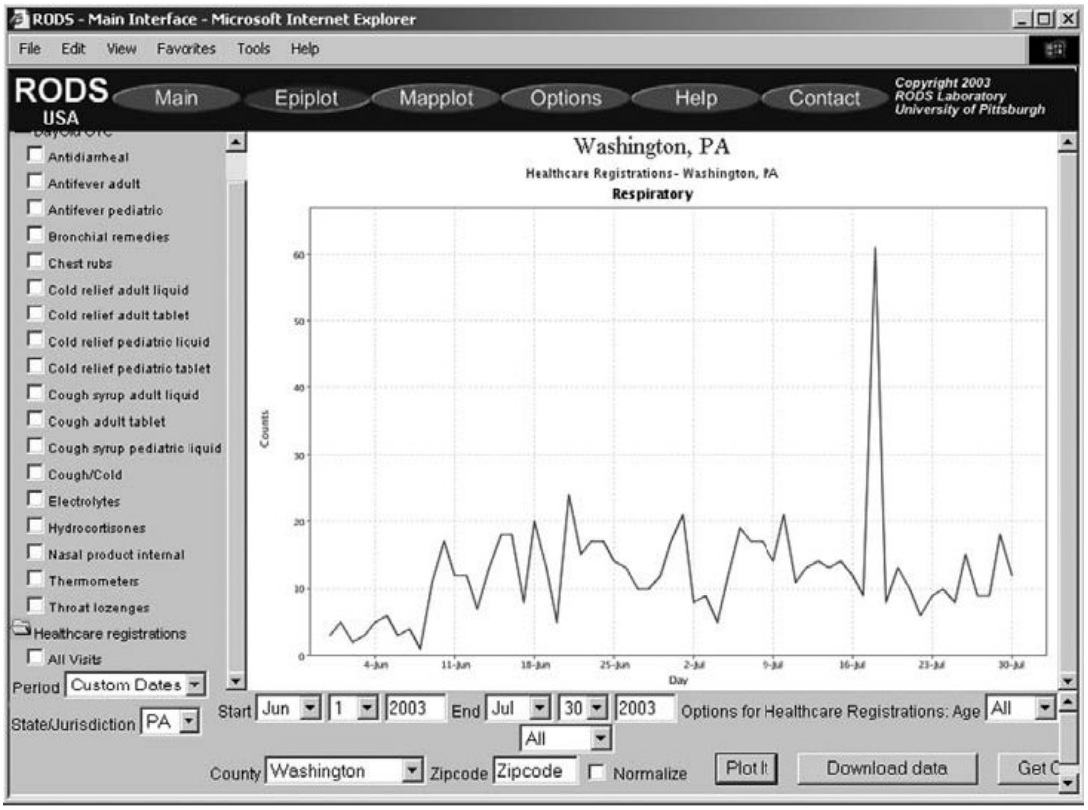

Figure 1-5: RODS Drill-down Screen
A drill-down screen from an earlier version of RODS.

1.2.2. BioSense

BioSense is a project by the Centers for Disease Control and Prevention (CDC), which was initiated in 2003 as a project to "enhance the nation's capability to rapidly detect, quantify, and localize public health emergencies, particularly biologic terrorism, by accessing and analyzing diagnostic and pre-diagnostic health data" (Loonsk, 2004). It collects and monitors LabCorp lab tests as well as Department of Defense and Department of Veterans Affairs diagnoses and procedures. It then provides some statistical analysis and visualization capabilities for public health officials to see and understand the data form their area. It currently supports 86 geographic regions (50 states, two territories, and 34 major metropolitan areas) (Sokolow et al., 2005). Its current mission is to "advance early detection by providing the standards, infrastructure, and data acquisition for near real-time reporting, analytic evaluation and implementation, and early event detection support for state and local public health officials."(Bradley et al., 2005) by attempting to provide a best-of-breed system for public health officials monitoring biosurveillance health series. In theory, its national scope and common interface could allow national collaboration and comparison across jurisdictions. But in 2006, CDC recognized that BioSense had not achieved the success that would be hoped for and began an analysis of performance to identify areas of improvement. Many practitioners use the system for data exploration rather than for the purpose of detecting outbreaks, due to the system's inflexibility and other limitations (Buehler et al., 2007). The CDC has since started an analysis and redesign of the system.

BioSense also incorporates EARS (Early Aberration Reporting System), which is an earlier CDC project designed to "provide national, state, and local health departments with several alternative aberration detection methods" (Hutwagner et al., 2003). It defines three aberration detection algorithms, which are often used as baseline algorithms for comparing new algorithms. These algorithms provide BioSense (and any other systems which care to use them) with basic aberration detection methods.

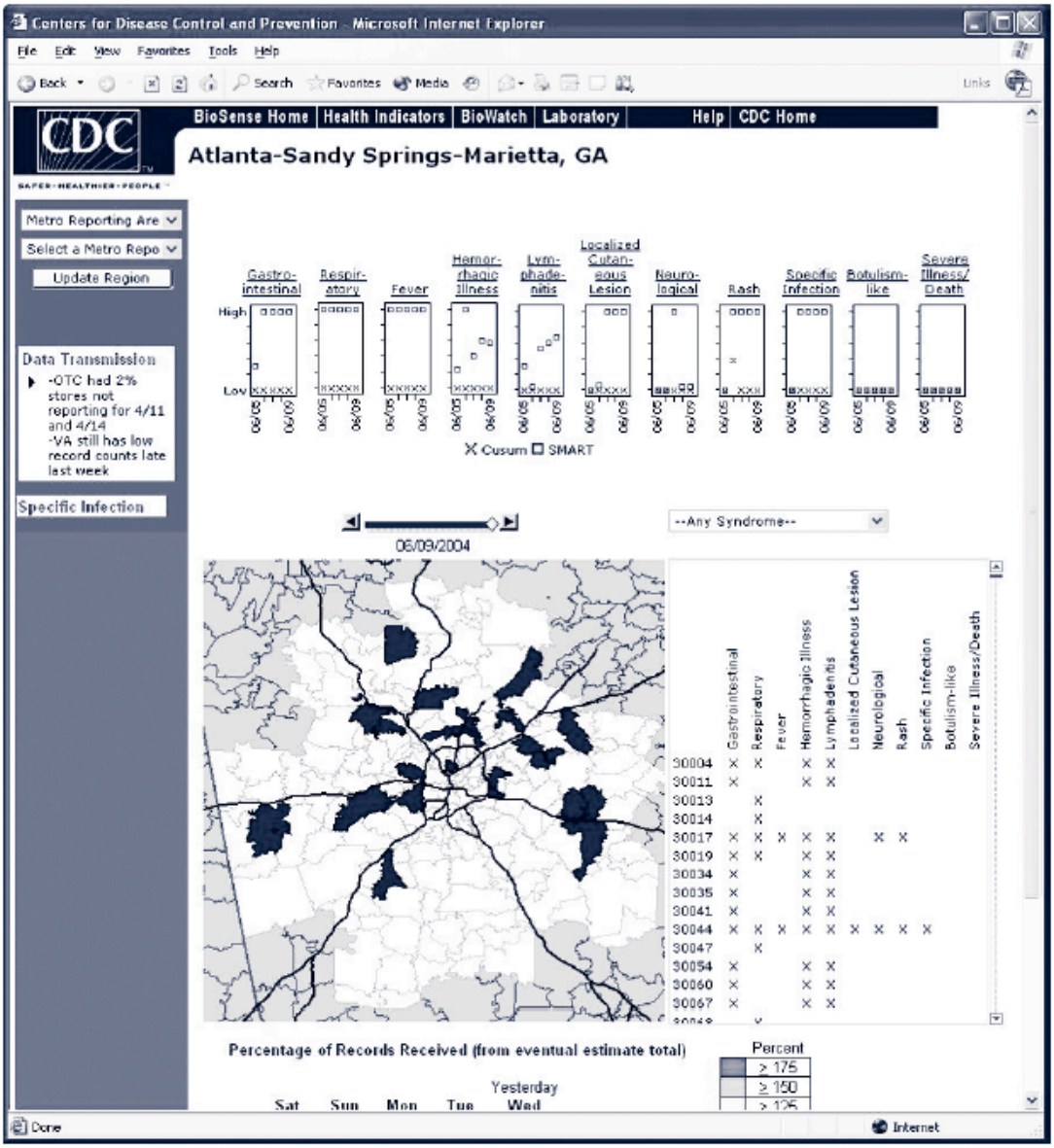

Figure 1-6: BioSense Screen Shot
BioSense example image, using demonstration data (from (Loonsk, 2004)]).

In general, the CDC is a main source of encouragement and support for biosurveillance research. It maintains a central website (CDC, 2006), a free online e-journal, and provides both tools and methodologies (such as BioSense and EARS) as well as funding for biosurveillance research. Its public implementations tend to be a few steps back from the cutting edge, but it provides invaluable support for biosurveillance research.

1.2.3. ESSENCE

ESSENCE (Electronic Surveillance System for the Early Notification of Community-Based Epidemics) is a collaboration between the Department of Defense Global Emerging Infections System and the Johns Hopkins University Applied Physics Laboratory (Lombardo et al., 2004). It uses data from military hospital visits, specifically diagnoses categorized into one of the International Classification of Diseases categories (ICD-9 codes), hospital site (identifying the hospital where the visit originated), patient's disposition (whether the record is for initial chief complaint, working diagnosis, or final diagnosis), and other data (age and gender of patient, clinic utilized, health care provider seen). It also includes "anonymized" consumer data, specifically hospital emergency room visits, physician office visits and over-the-counter drug sales. It then provides visualization and analysis of those data. This is mainly done for DoD use, but several of the methods developed for ESSENCE have been published in the scientific literature and it is also used by epidemiologists in the Washington, D.C. area.

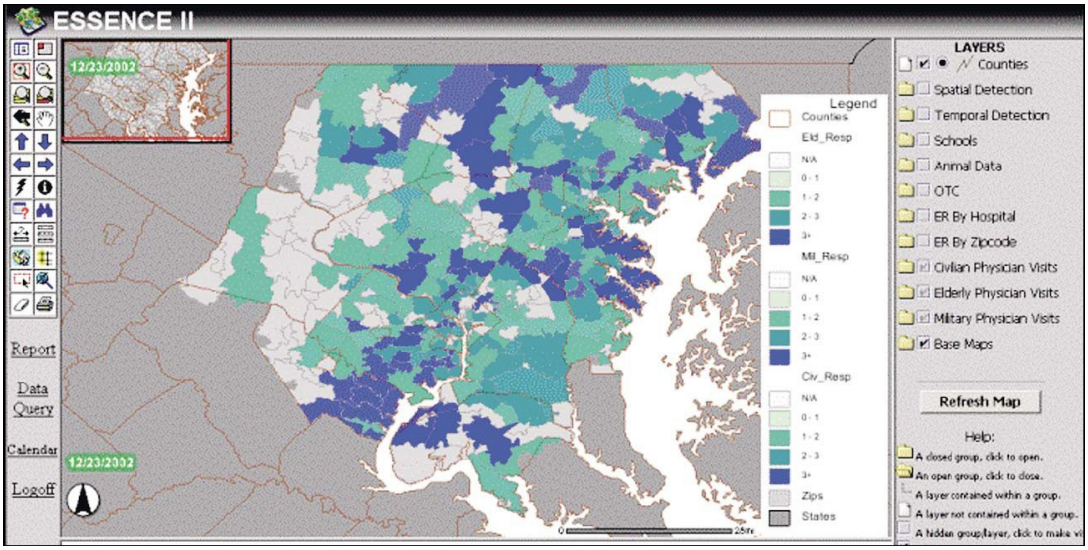

Figure 1-7: ESSENCE Screen Shot
An ESSENCE screen shot showing incidence of respiratory counts in the National Capital Area based on military and civilian physician visits.

1.2.4. Other Systems and Systems Proposals

While the three systems described above (BioSense, RODS, and ESSENCE) are the largest and most significant, many other city and state public health departments have developed their own systems. Most of these areas use similar methods taken from current research or larger systems, but two deserve special attention: the Olympics monitoring systems and New York City's public health monitoring.

The Olympics are an excellent test case for biosurveillance systems. The Olympic city has a diverse population, tightly packed, with peak athletic performances on the line. Recent Olympics have developed biosurveillance systems to detect any disease spread, either using unique systems (Dafni et al., 2004) or based on existing technology such as RODS (Gesteland et al., 2003).

New York City is both the largest city in the U.S. and one of the most visible targets for terrorists. It is only natural that it would also have the largest public health monitoring system. A history of the system is given by (Heffernan et al., 2004); in 1995 it began to monitor diarrheal illness at nursing homes, surveillance of stool submissions at clinical laboratories, and over-the- counter (OTC) pharmacy sales for diarrheal illness. It later grew to include prescription drug sales, ER visits, and worker absenteeism. Recent presentations have shown the evolving NYC system, growing to include spatial scan statistics (Mostashari, 2002) as well as multivariate combinations and visualization techniques (Paladini, 2006).

Any new system requires a number of components. A number of researchers and public health officials have attempted to define what would be necessary components of a biosurveillance system (Bean & Martin, 2001, Wagner et al., 2003, Pavlin et al., 2003). While most of these definitions have been supplanted by analyses of and reactions to actual systems (such as (Buehler et al., 2007)), they still provide a fairly comprehensive view of what is involved in creating a new biosurveillance system. When considering the creation of a new system, one cannot consider only the algorithms used (which we analyze in this dissertation) but must also consider larger issues such as data collection and privacy concerns. While the algorithms we present should improve such systems, we reiterate that a real system involves much more than the detection component we focus on here.

1.3. Data Sets Used in this Dissertation

In examining and testing the ideas in this dissertation, we use four main sources of biosurveillance data. These were used to compare the effectiveness of different methods, to test the validity of assumptions, and to find appropriate parameters. By using authentic biosurveillance data, we can be more confident that the ideas presented here are valid and practical in real-life scenarios.

1.3.1. BioALIRT

Our first authentic data set comes from the BioALIRT program conducted by the U.S. Defense Advanced Research Projects Agency (DARPA), described in (Siegrist & Pavlin, 2004). Permission to use the data was obtained through data use agreement #189 from TRICARE Management Activity. The data set includes three types of daily counts: military clinic visit diagnoses, filled military prescriptions, and civilian physician office visits. The BioALIRT program categorized the records from each data type as respiratory (Resp), gastrointestinal (GI), or other. The data were gathered from 10 U.S. metropolitan areas with substantial representation of each data type. The data consist of counts from 700 days, from July 1, 2001 to May 31, 2003. As an example, we use the daily count of respiratory symptoms from civilian physician office visits, all within a particular U.S. city (cities are not identified, due to privacy concerns), which can be seen in Figure 1-8. The same series is displayed in Figure 1-9, which shows the data split by day-of-week. In this, you can see the weekend/weekday difference much more clearly.

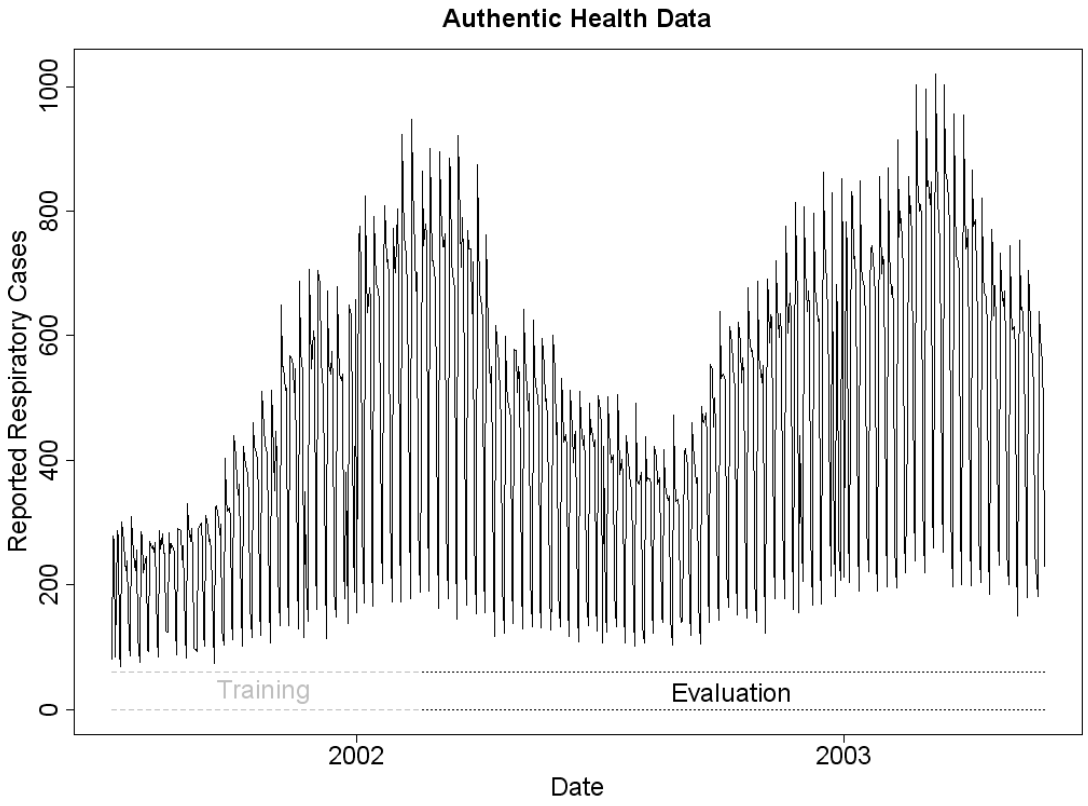

Figure 1-8: BioALIRT Respiratory Data Example
Daily counts for reported respiratory symptoms among civilians, from the BioALIRT data set. The first 1/3 of the data (233 days) will generally be used for training, and the last 2/3 (467 days) for evaluation.

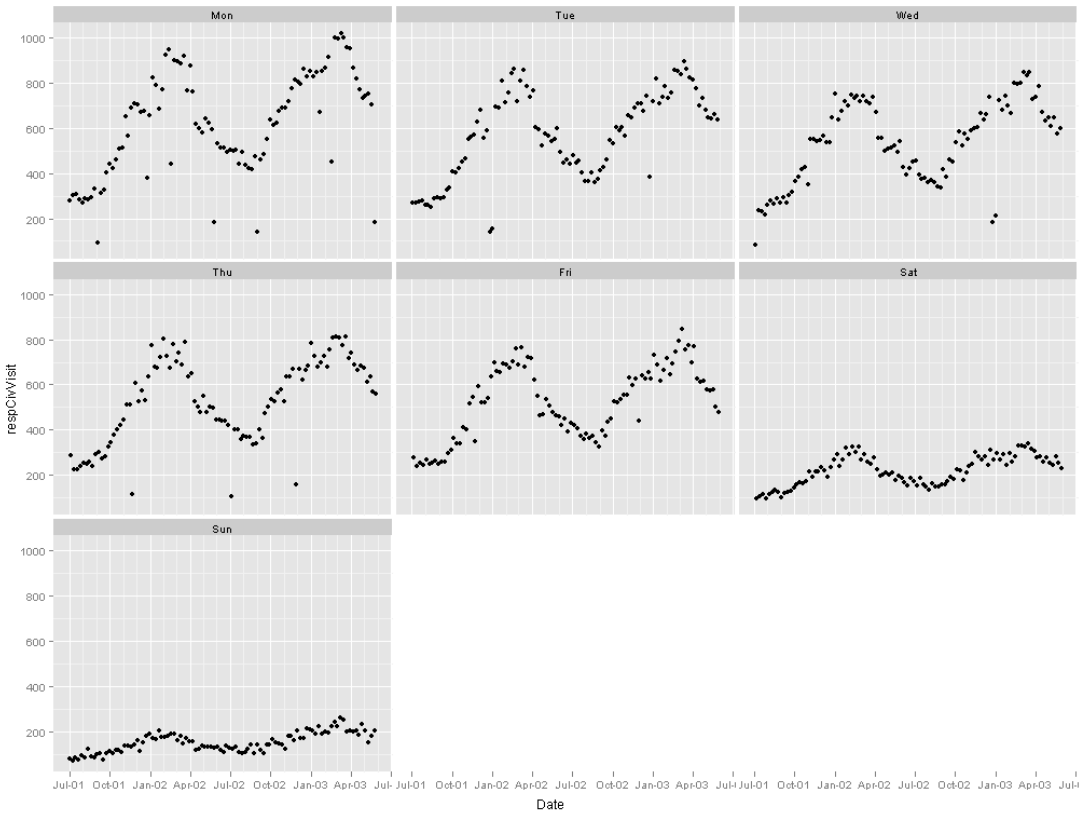

Figure 1-9: Seasonal Subseries Plot for BioALIRT Respiratory Data
Daily counts for respiratory symptoms among civilians, from the BioALIRT data set, split by day-of-week.

1.3.2. Over-the-counter (OTC) medication sales

The second data set comes from a grocery chain in the Pittsburgh area. It includes daily sales for eight categories of medications, from August 1999 to January 2001 (Goldenberg et al., 2002a). The eight data streams are

Asthmatic remedies (Asthmatic.Remedies),

Allergy medicine (Allergies.Caps),

Cough syrups/liquid decongestants (Cough.Syr.Liquid.Decongest),

Nasal sprays (Nasal.spray.drops.inhalar),

Non-liquid decongestants (room.decongest),

Pills (tabs.caps),

Time release pills (tabs.caps.time.release), and

Throat lozenges/cough drops (throat.loz.cough.drops).

A set of charts that include a timeplot, zoomed time plot, autocorrelation function (acf) plot, and quantile-quantile (Q-Q) normal plot is shown for three of the series in Figure 1-10.

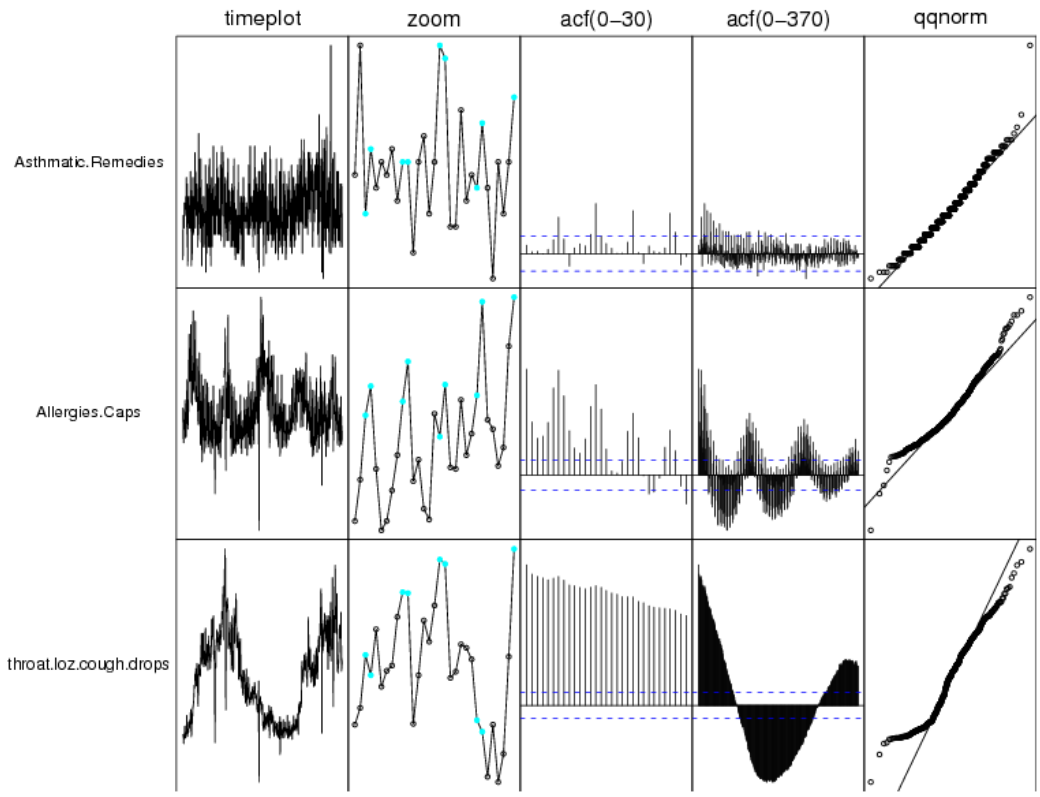

Figure 1-10: OTC Series Summary Visualizations
Summary graphs for three OTC categories. Average daily counts vary largely across different categories, with varying degrees of weekly and annual dependence.

1.3.3. Chief complaints at emergency departments

The third data set, from ESSENCE (Electronic Surveillance System for the Early Notification of Community-Based Epidemics), is composed of 35 time series representing daily counts of ICD-9 codes. ICD-9 is the 9th edition of the

International Statistical Classification of Disease and Related Health Problems, published by the World Health Organization (WHO) and used worldwide. It describes a set of ICD-9 codes in order to standardize classification of a wide variety of health conditions, mainly symptoms and diseases. Our data set consists of ICD-9 codes generated by patient arrivals at emergency departments (ED) in an unspecified metropolitan region from Feb-28-1994 to Dec-30-1997. The 35 series were then grouped into 13 series, using the CDC's syndrome groupings.

These syndrome groups show the diversity across the different syndrome subgroups in the level of daily counts and in weekly and annual dependence. We removed the counts for the 38 holidays contained in the data set, as their values are significantly different from non-holidays, and holidays will occur on predictable dates in the future. In the following we use three series for display (Gastrointestinal (GI)-related, Respiratory (Resp), and Unexplained Death (Unexpl Death) ED visits). These are shown in Figure 1-11.

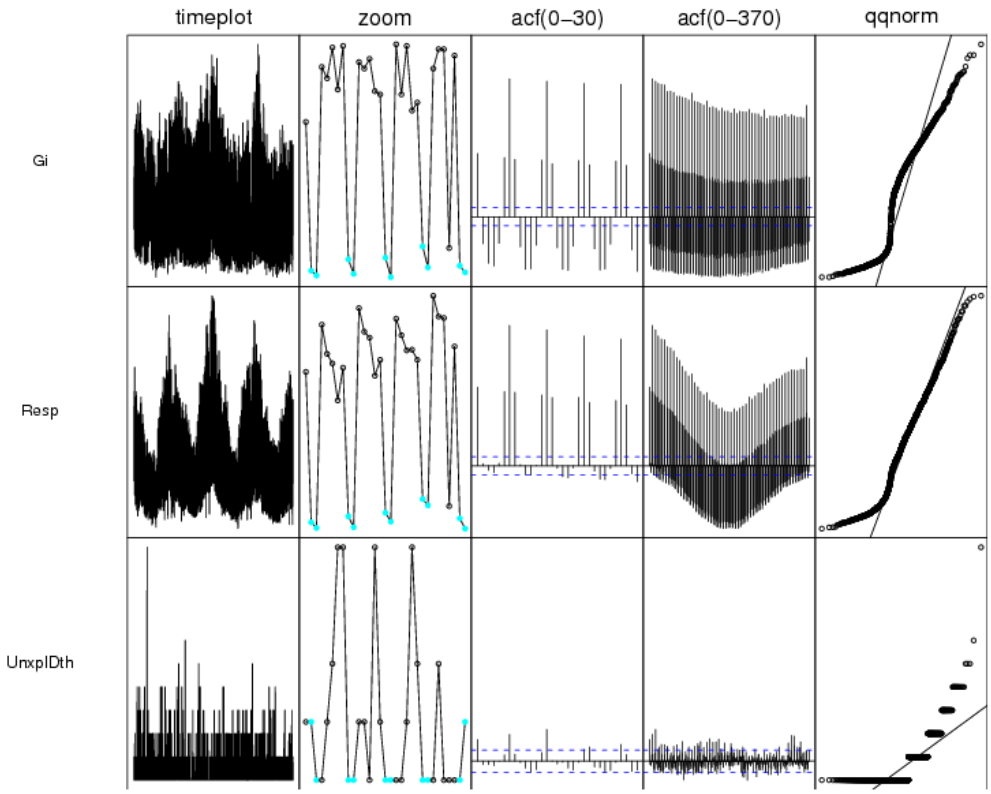

Figure 1-11: ED Series Summary Visualizations
Summary graphs for three ED categories. Low-count series like UnexplDth bring additional challenges to biosurveillance monitoring.

1.3.4. ISDS contest data

In 2007, the International Society for Disease Surveillance (ISDS) organized a technical contest. Participants were "encouraged to develop novel techniques or test state-of-the-art alerting algorithms for prospective disease outbreak detection on realistic data." In order to do this, surveillance data sets were provided by the Canadian Network for Public Health Intelligence (CNPHI), which agreed to make them permanently available for academic use after the contest. The contest used three types of data:

1. Patient emergency room visits (ED) with gastrointestinal symptoms
2. Aggregated over-the-counter (OTC) anti-diarrheal and anti-nauseant sales

3. Nurse advice hotline calls (TH) with respiratory symptoms

These data were based on a three-year historical data set from Winnipeg, Manitoba, Canada with a population size just over 700,000. This data set was used to model the characteristics and trends present in the contest baseline data. In addition, three types of outbreaks were simulated and inserted. The contest outbreak profiles were modeled after data effects of three historical outbreaks, each affecting a single data type. From the contest description:

1. *In the spring of 2000, the community of Walkerton, Ontario experienced one of the worst outbreaks of waterborne E.coli 0157:H7 in Canadian history. ED data for gastrointestinal (GI) symptoms retrospectively collected from the local hospital clearly showed the outbreak profile.*
2. *A similarly large waterborne outbreak of Cryptosporidium occurred in the Battleford area of Saskatchewan during the spring of 2001. Due to the prolonged, less severe nature of Cryptosporidium, many infected residents self-medicated, evidenced by an increase of OTC anti-diarrheal and anti-nauseant product sales during the outbreak.*
3. *Large-scale, seasonal influenza epidemics (such as bird flu) have not been widely characterized through syndromic surveillance systems. Because nurse hotlines are commonly used by residents to report symptoms of influenza like illness in the Winnipeg region, this data stream was chosen for this outbreak. The profile is a combination of the few historical examples available in publication.*

Each data type had thirty 'scenarios', which consisted of the same baseline data with a different stochastically generated outbreak inserted. Each data type had five years of data, and the outbreak was inserted somewhere in the last four years. An example of stochastic outbreaks is seen in Figure 1-12, which shows an exemplar outbreak (for the influenza outbreak injected into the nurse hotline/TH series) as well as thirty stochastic instances of actual outbreak counts (seen as thinner colored lines).

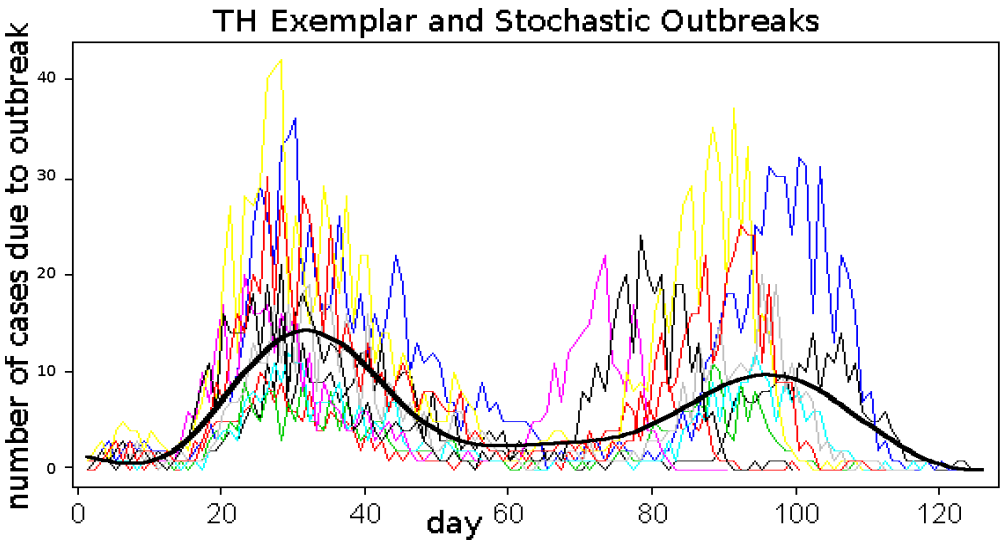

Figure 1-12: ISDS Contest Exemplar and Simulated Stochastic Outbreaks
Exemplar influenza outbreak inserted into nurse hotline calls (thick black line) and stochastic instances of the same (thin colored lines).

1.4. Existing Research on Statistical Methods for Biosurveillance

1.4.1. Control Chart Methods

Statistical control charts, invented by Walter Shewhart and used as the basis of Statistical Process Control (SPC), were first used in the 1920s to monitor factory outputs to discover abnormally high rates of product defects. An alarm indicated variance beyond the normal operating conditions and the presence of a "special cause", which was usually a faulty process that could then be corrected. Control

charts are statistical tools for monitoring process parameters and alerting when there is an indication that those parameters have changed. They are now widely used in health-related fields, particularly in biosurveillance (as seen in (Benneyan, 1998b, Woodall, 2006)). There are some difficulties in directly applying control charts to daily pre-diagnostic data, since classical control charts assume that observations are independent, identically distributed, and typically normally distributed (or with a known parametric distribution). However, as described in Section 1.4.2, such assumptions generally do not hold for the pre-diagnostic data being considered.

Control charts are usually two-sided, monitoring for an increase or decrease in the parameter of interest. Monitoring is done using an upper control limit (UCL) and lower control limit (LCL), respectively. In biosurveillance, we are usually only concerned with a significant *increase* in the underlying behavior indicative of a disease outbreak, and therefore only a UCL is used. The control chart is applied to a sample statistic (often the individual daily count), and alerts when that statistic exceeds the UCL. This UCL is a constant, set to achieve a certain false alert level; the true alert rate can then be computed.

The three main types of control charts are the Shewhart, Cumulative Sum (CuSum), and Exponentially Weighted Moving Average (EWMA). These are covered in detail in (Montgomery, 2001), but we provide a basic description here:

Shewhart. The Shewhart chart is the most basic control chart. A daily sample statistic (such as a mean, proportion, or count) is compared against upper and/or

lower control limits (UCL and LCL), and if the limit(s) are exceeded, an alarm is raised. The control limits are typically set as a multiple of standard deviations of the statistic from the target value (Montgomery, 2001). It is most efficient at detecting medium to large spike-type outbreaks.

CuSum. Cumulative-Sum (CuSum) control charts monitor cumulative sums of the deviations of the sample statistic from the target value. CuSum is known to be efficient in detecting small step-function type changes in the target value (Box & Luceno, 1997).

EWMA. The Exponentially Weighted Moving Average (EWMA) chart monitors a weighted average of the sample statistics with exponentially decaying weights (NIST, 2004). It is most efficient at detecting exponential changes in the target value and is widely used for detecting small sustainable changes in the target value.

The classic Shewhart chart for monitoring the process mean relies on drawing a sample from the process at some frequency (e.g., weekly), and plotting the sample mean on the chart. CuSum and EWMA are similar, except that the plotted value is a more complex function of the current and previous samples. Parameter limits are defined such that if the process remains in control, nearly all of the sample means will fall within the control limits. If a sample mean exceeds the control limits, it indicates that the process mean has shifted, or in other words, the process has gone out of control; an alarm is triggered and an investigation follows to find its cause(s) (Page,

1954, Reinke, 1991). Figure 1-13 shows an example of a one-sided Shewhart control chart on simulated random data, for detecting increases in the process mean. The dotted line indicates the control limit; red stars show points exceeding the limit.

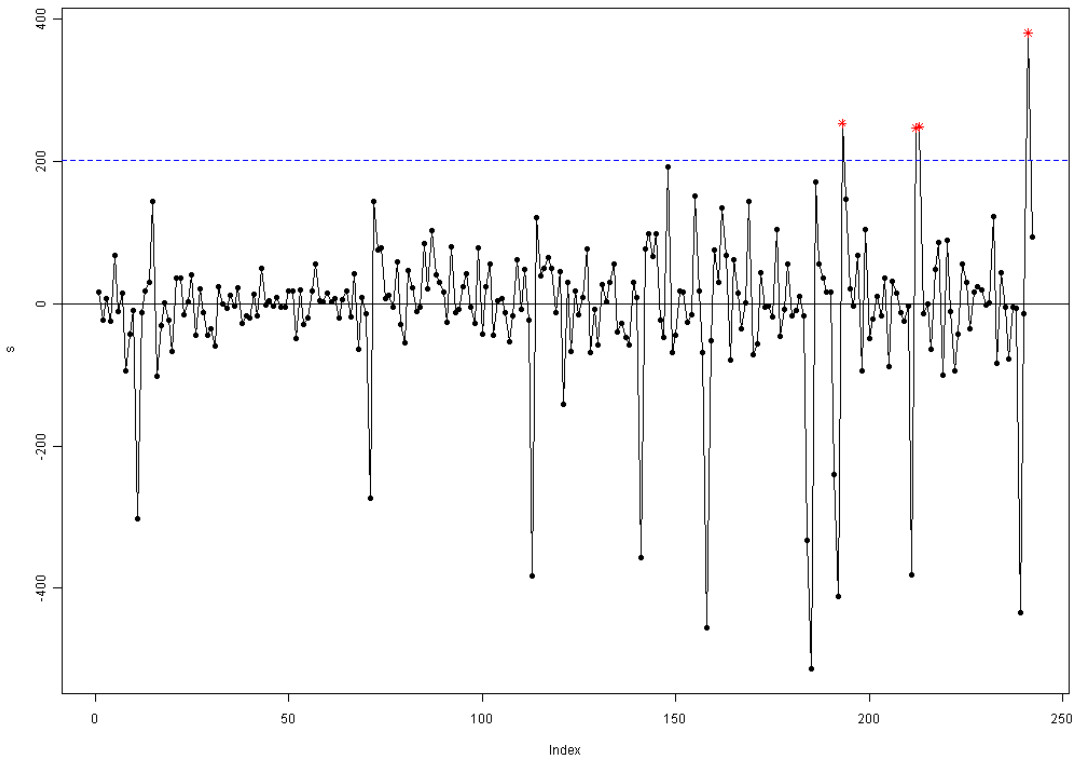

Figure 1-13: Shewhart Control Chart
Sample Shewhart Control Chart. The dashed blue line is the control limit; red stars are points exceeding the control limit.

Table 1-1 summarizes for each of the three charts the monitoring statistic (denoted Shewhart_t, EWMA_t and CuSum_t), the upper control limit (UCL) for alerting, the parameter value that yields a theoretical 5% false alert rate, and a binary output indicator that indicates whether an alert was triggered on day t (1) or not (0). Let Y_t denote the raw daily count on day t. We consider one-sided control charts where an alert is triggered only when there is indication of an increase in mean (i.e., when the

monitoring statistic exceeds the UCL). This is because only increases are meaningful in the context of health care seeking counts.

	Shewhart	**EWMA**	**CuSum**
Monitored Statistic	$Shewhart_t=Y_t$	$EWMA_t=$ $\lambda Y_t+ (1-\lambda)EWMA_{t-1}$	$CuSum_t=max(0,$ $CuSum_{t-1}+Y_t- \sigma/2)$
UCL	$UCL=\mu+k\sigma$	$UCL=EWMA_0+k\sigma$ $s^2=\lambda/(2-\lambda)\sigma^2$	$UCL= \mu+h\sigma$
Output	$S_t=$ if $[Shewhart_t>UCL]$	$E_t=$if $[EWMA_t>UCL]$	$C_t=$if $[CuSum_t>UCL]$

Table 1-1: Features of three main control charts

One point to remember is that in biosurveillance, the CuSum and EWMA are "reset" after an alert. In other words, after an alert, the statistic is re-initialized (usually to 0, though variants include setting the statistic to the mean observed value or the last observed value before the alert). This is done because the false alert rate determines the amount of resources which must be devoted to a system. Resetting ensures that the ATFS is both the average time to first false signal and the average time between false signals; thus the overall false alert rate will be 1/ATFS, even though the rate will not be constant for each day.

(Reinke, 1991) was one of the first to suggest the use of industrial SPC techniques for prospective epidemiologic investigations; he describes both a regression method for normalization and a negative binomial Shewhart chart for detecting outbreaks. Soon after, (Hutwagner et al., 1997) used a slight modification of the CuSum for detecting Salmonella outbreaks. In subsequent years, others such as (Radaelli, 1992) have used such techniques as the CuSum for detecting rare events, as it can provide increased sensitivity for small outbreaks occurring over a period of time (a position recently supported by (Fricker et al., 2008b)). With the growth of biosurveillance in the late

1990's, SPC methods became increasingly used in hospitals (Benneyan, 1998a, Benneyan, 1998b) as well as for epidemiologic disease surveillance (Farrington et al., 1996). Most of the systems in practice use SPC as the main detection component. BioSense uses CuSum at the state level (Bradley et al., 2005), EARS provides three Shewhart-based methods (with different sliding windows for the estimated baseline) (Hutwagner et al., 2003). RODS (Tsui et al., 2003) and ESSENCE (Marsden-Haug et al., 2007) also use SPC methods. Some research (and systems such as ESSENCE) use distributions other than normal, such as Poisson (Rogerson & Yamada, 2004) or negative binomial (Reinke, 1991). Over the last several years, SPC has become the standard method rather than an exception (Woodall, 2006).

1.4.2. Biosurveillance Surveys and Challenges with Biosurveillance data

A number of articles have described the various problems with analyzing biosurveillance data. These include (Burkom, 2003b, Fienberg & Shmueli, 2005, Fricker & Rolka, 2006, Shmueli & Burkom, 2009) and several others, usually in conjunction with a review of the approaches used to tackle those problems. The problems include inherent noise in pre-diagnostic data, which provides no firm conclusion of a specific disease but provides total counts of symptoms which can come from a variety of diseases; the fact that a variety of diseases or even non-diseases such as holidays, celebrity diseases, or weather can influence the counts; the non-stationarity of the time series, which vary both over the long term and in the shorter terms of annual or weekly patterns; the autocorrelation inherent in the health series; the non-normality of the data; and the lack of standards for identifying outbreaks and testing algorithms. These problems cause particular issues for control

chart detection methods which assume well behaved normal iid data. There have also been a host of reviews of biosurveillance research. Buckeridge (Buckeridge et al., 2004, Buckeridge et al., 2005, Buckeridge, 2007, Buckeridge et al., 2008) continues to periodically analyze the state of the art, but many others also provide surveys of existing methods (Bravata et al., 2004, Farrington & Andrews, 2004, Reingold, 2003, Rolka, 2006, Sonesson & Bock, 2003, Wagner et al., 2001).

1.4.3. Preprocessing Methods

As in the industrial setting, control charts are used to monitor time series data to detect "special causes" or abnormalities; in this case, such abnormalities are potentially indicative of an outbreak. However, currently collected biosurveillance data violate most of the assumptions required of data monitored by control charts. Underlying all of the SPC methods is the assumption that the monitoring statistics are independent and identically distributed (iid), with the distribution generally assumed normal (although modifications can be made for statistics with known, non-normal distribution). While control charts are very effective for monitoring processes that meet the independence and known distribution assumptions, they are not robust when these assumptions are violated (Shmueli & Fienberg, 2006). Thus, alarms triggered by control charts applied directly to raw syndromic data can arise not from actual outbreaks but due to explainable patterns in the data. Reports of very high false alarm rates from users of current syndromic systems lend evidence to this claim.

The explainable patterns are caused by factors unrelated to a disease. As an example, it is quite common for doctors' offices to have reduced staffing on weekends.

Therefore, data on daily doctor visits will see an explainable and predictable drop on Sundays and a corresponding increase on Monday. Many syndromic data streams demonstrate a marked day-of-week (DOW) effect, dropping or increasing in counts over the weekends, with an early work-week resurgence or drop. Holidays and other external factors can cause a similar phenomenon. Even the release of Harry Potter books has a measurable effect on hospital admissions (Gwilym et al., 2005).

If the control chart assumptions do not hold, the charts will fail to detect special cause variations and/or they will alert frequently even in the absence of special cause variations. Therefore, much research has attempted to preprocess the health data by forecasting the expected level and monitoring the residuals. Many different techniques have been proposed to forecast the health data, with varying degrees of success. In the following we describe the main methods used for predicting next-day counts. We denote by Y_t the count on day t, and by f_t the forecasted count for day t.

Regression models are the most popular method for forecasting daily health series counts. In this case, several time-variant predictors are assumed to combine linearly to produce the expected level of health activity on a given day. More formally, the daily counts are modeled as:

$$Y_t = \beta_{0t} + \beta_{1t}x_{1t} + \beta_{2t}x_{2t} + ... + \epsilon_t \qquad \text{(Eq. 1-1)}$$

where each ϵ_t is an independent identically distributed normal variable, $\epsilon_t \sim N(0, \sigma^2)$ and the model parameters β are estimated by least squares. Predicted counts are then calculated using

$$f_t = E(\hat{y_t}) = \hat{\beta}_{0t} + \hat{\beta}_{1t}x_{1t} + \hat{\beta}_{2t}x_{2t} + ... \quad \text{(Eq. 1-2)}$$

A number of variations on the basic regression model have also been used. In particular, the choice of predictors varies. Serfling (Serfling, 1963) proposed a way of incorporating annual seasonal patterns by using sine and cosine predictors with a period of 365.25 days, e.g., $\sin(\frac{2\pi}{365.25}t)$. While this was proposed for retrospective analysis of pneumonia incidence, it can be used for prospective modeling as well, for any series which follows a roughly sinusoidal annual pattern. Day-of-week dummy variables ($x_{Mon}, ..., x_{Sat}$) are common, as is a linear trend term (t) (as in (Brillman et al., 2005)). A dummy variable for holidays and day-after-holidays is also sometimes used, although the holiday effect does not always follow official holidays (as seen in (Kikuchi et al., 2007)). Non-linear regressions such as Poisson regression, or linear regression of $\log(y_t)$ rather than y_t are also used, under the assumption that the predictors used have a multiplicative rather than additive effect on counts (such as in (Kleinman et al., 2004)). Regression forecasting is used in some variant by nearly all existing biosurveillance systems; for example, BioSense uses SMART scores (a type of Poisson regression) at the zip code level.

7-day differencing, as proposed in (Muscatello, 2004), is perhaps the simplest forecasting model. It models the next day's expected count as the count from the same day of week, one week earlier: $f_t = Y_{t-7}$.

Exponential Smoothing is a method, originally developed in the 1950's by Brown (Brown, 1959) and others, which uses a weighted sum of past observations to predict the next observation, where the weights are exponentially decaying over time. The forecast is given by

$$f_t = \lambda \sum_{i=1}^{t-1} Y_i (1-\lambda)^{t-1-i} \qquad \text{(Eq. 1-3)}$$

where λ is a smoothing parameter between 0 and 1, that determines the weight given to recent observations. The forecast is easily computed as

$$f_t = Y_{t-1}\lambda + f_{t-1}(1-\lambda) \qquad \text{(Eq. 1-4)}$$

Its statistical properties are discussed further in (Chatfield et al., 2001).

ARIMA (AutoRegressive Integrated Moving Average) models are statistical time series models for analyzing and forecasting time series data. While they have not often been used in biosurveillance (an exception is (Reis & Mandl, 2003) and more recently, (Shtatland et al., 2009)) due to their complexity of implementation and difficulty of automation, they seem to be a reasonable method when employed.

Holt-Winters multiplicative exponential smoothing (Chatfield, 1978) is a recently adopted method which captures level, trend, and day-of-week effect and smoothly changes its parameters over time. In addition to being easy to understand and implement for a large class of data types, it has been shown (Burkom et al., 2007) that this method is very effective in the context of biosurveillance. Little data history is needed, and due to its highly adaptive nature,

it reduces the need for individual modifications for specific data sources and syndrome groupings. The Holt-Winters method is discussed further in Section 3.2.3.

1.4.4. Other Detection Methods

There have also been a variety of more unusual methods proposed for detection of disease outbreaks. These include methods adopted from machine learning, such as the neural network approach in (Adams et al., 2006). A review of biosurveillance ideas from data mining was presented in (Moore et al., 2002). Some techniques come from other disciplines, such as the use of wavelets for describing a time series in chemical process control (Shmueli, 2005, Stacey et al., 2005).

Some approaches consider monitoring deviations other than an increase above the expected level. (Nobre & Stroup, 1994) use exponential smoothing to forecast the next-day count, but monitor the differences in the first derivative to see if the rate of increase is larger than expected. The moving-F statistic proposed by (Riffenburgh & Cummins, 2006) looks for a change in variance. (Naus & Wallenstein, 2006) look at adapting the spatio-temporal scan statistic to a purely temporal detection method.

Bayesian approaches are also gaining prominence. Wong (Wong et al., 2002, Wong et al., 2003a, Wong et al., 2003b, Wong, 2004) suggests using a Bayesian analysis over multiple subsets of data (both temporal and geographical) to detect recent events of interest, a method which has been incorporated into RODS. One of the most promising new approaches (described in (Ozonoff & Sebastiani, 2006, Martinez-

Beneito et al., 2008)) uses a Bayesian model with two different settings: zero when there is no outbreak, and one when there is an outbreak. This allows the estimation of the likelihood of an outbreak, as well as a sense of its posterior distribution for the current day.

1.4.5. Data Sources and Multivariate Detection

The question of which data sources to use is also a recurring one. Most early studies use correlation between health data sources and the disease of interest as a way of indicating the usefulness of a data source. This includes over-the-counter electrolyte sales (Hogan et al., 2003); over-the-counter medications (Goldenberg et al., 2002a); blood donor screenings (Kaplan et al., 2003); preliminary laboratory tests (Najmi & Magruder, 2004, Widdowson et al., 2003); and using influenza-related Internet search terms (Polgreen et al., 2008). A recent study investigates the predictive value of various case definitions (Guasticchi et al., 2008) and attempts to compare the performance of various data sources for detecting specific diseases. Similarly, the recently announced Google approach (Ginsberg et al., 2009) attempts to automatically find a good combination of search terms which leads to maximum predictive value.

Recently, the issue of multivariate data streams has been the target of growing attention from the CDC and other researchers (Shmueli & Fienberg, 2006). The challenges of biosurveillance are too significant to not take advantage of all available information, and the fact that there are generally multiple health data streams which can be monitored within a specific geographical area means that they have the

potential to gain more information about the indicators of an outbreak. Some research has shown that monitoring multiple data streams can result in a detection improvement over univariate monitoring (Lau et al., 2008). This research includes the process of selecting which data sources to use (Mandl et al., 2004) as well as determining what the circumstances are for performing different types of multivariate alerting combinations (Burkom et al., 2005). Others have performed research into directionally sensitive versions of multivariate detection algorithms (Fricker, 2006, Yahav & Shmueli, 2007). When the multivariate reports are hierarchical, the consideration of this hierarchy and its aggregation or disaggregation can also have an effect on performance (Burkom et al., 2004). Special consideration is also given when the multivariate time series come from different locations, rather than being measures of different syndromes within the same location (Hong & Hardin, 2005). Finally, multivariate data can be used to improve forecasting methods; Najmi and Magruder (Najmi & Magruder, 2005) used multichannel least-mean-squares (LMS) and Finite Impulse Response (FIR) filters, with a recursive fitting algorithm, to improve forecasting performance.

1.4.6. Performance Comparison

In order to determine which algorithm is most effective at detecting disease outbreaks, one must compare the detection algorithms in a reasonable way. Most individual studies use personal data sets and often do not provide comparisons against other algorithms. In evaluating, most authors use a system of inserting simulated outbreaks into authentic historical health data and use metrics analogous to those described in Section 1.1.2 (Hutwagner et al., 2005b, Kleinman & Abrams, 2006,

Stoto et al., 2006, Wallstrom et al., 2005). Some researchers (such as (Reis et al., 2003)) evaluate performance slightly differently, by judging detection on a per-day basis rather than a per-outbreak basis. Instead of determining how many outbreaks were detected, they measure the proportion of outbreak days on which the algorithm alerted. However, we believe that the purpose of a biosurveillance system is more directly measured by how many outbreaks it detects; providing an extra alert during one outbreak is less useful than detecting an additional outbreak.

Only recently have there been evaluations attempting to determine what causes different algorithms to perform better. (Burkom & Murphy, 2007b) analyzed the effect of different types of data series on different algorithm performance. (Fricker et al., 2008b) conducted a study comparing CUSUM methods against EARS, then delved further into comparing CUSUM and Shewhart detection methods on different outbreak types. (Buckeridge et al., 2008) went even further, analyzing the EARS methods on the basis of their underlying algorithm qualities (inclusion of a guard band and use of previous days' data) to discern the effects on detection performance. This sort of analysis represents a growing sophistication in algorithm comparison, determining not only which algorithms perform better, but why.

Two competitions have been held in an attempt to compare algorithms' performance against each other. The first was the BioALIRT challenge in 2004, which compared different teams' performance in detecting fifteen outbreaks identified by experts over data from five American cities (Siegrist & Pavlin, 2004, Siegrist et al., 2005). Some

competitors questioned the accuracy of the labeling of those outbreaks, but the competition was very successful in bringing together numerous different biosurveillance research teams and comparing their performance on a common problem. More recently, the ISDS (International Society for Disease Surveillance) hosted a competition (Burkom, 2010), using data based on Canadian health disease outbreaks. As more data sets become publicly available (either authentic data or realistic simulated data as described in Section 1.4.7), such comparisons between algorithms will become easier to perform. By comparing algorithms on the same data, over a wider variety of data, the relative performance of algorithms on different types of data and outbreaks will become clearer.

Some studies have also been done to test the effectiveness of actual detection systems at providing early detection of outbreaks; these studies can be especially valuable, as they can provide 'gold standard' data, with days labeled as outbreaks by actual health professionals. (Hope et al., 2008a) performed such an evaluation with an Australian biosurveillance system and with the biosurveillance detection after natural disasters (Hope et al., 2008b). Effectiveness tests have also been performed by others on influenza (Lee et al., 2002) and an overseas U.S. armed forces system (Meynard et al., 2008). Still, biosurveillance has thus far also failed to live up to the promise of a strong detection system with few false alerts, and this failure has been mentioned in reviews by (Stoto et al., 2004) and others, as well as by (Sullivan, 2003). Sullivan also suggested that combining pre-diagnostic data with biosensors for disease agents would be more effective. Over the last five years, research has improved the

sensitivity and specificity of biosurveillance methods, but there is still progress which needs to be made in order for biosurveillance systems to show value for early detection.

1.4.7. Simulating Health Series

A major barrier to evaluating surveillance algorithms has been data accessibility: typically researchers do not have access to biosurveillance data unless they are part of a biosurveillance group. This means that a very limited community of academic researchers works in the field, with a nearly impenetrable barrier to entering it (especially for statisticians or other non-medical academics). Furthermore, different research groups use different privately held data to test their detection algorithms, often based on existing agreements with associated local organizations. For example, Pittsburgh-based researchers at Carnegie Mellon University and the University of Pittsburgh use Emergency Department and over-the-counter drug sales data for Allegheny County, Pennsylvania (Neill et al., 2005), but the Australian Centre for Epidemiology and Research uses influenza cases in New South Wales (NSW), Australia (Zheng et al., 2007). The confinement of each research group to a small and limited set of data and the lack of data sharing across groups "leaves opportunity for scientific confounding" (Rolka, 2006). In other words, it makes it uncertain whether the difference in results is due to the difference in algorithm or the difference in data.

One way to address this problem is to generate simulated data sets which can be freely used by different groups of researchers. While simulated data have their own

difficulties, they seem to be a necessity for modern biosurveillance research. (Buckeridge et al., 2005) explain that "[they] are appealing for algorithm evaluation because they allow exact specification of the outbreak signal, perfect knowledge of the outbreak onset, and evaluators can create large amounts of test data." In order to be useful, of course, they must have the same characteristics as authentic data. In Section 5.2, we describe a method for evaluating simulated data on its similarity to authentic health data. Here, we describe the simulation methods which have been proposed for use in biosurveillance.

The first implementation of wholly simulated biosurveillance data in the form of daily counts is the publicly available simulated background and outbreak data sets by (Hutwagner et al., 2005a). The background series are generated from a Negative-Binomial distribution with parameters set such that "Means and standard deviations were based on observed values from national and local public health systems and biosurveillance surveillance systems. Adjustments were made for days of the week, holidays, post-holiday periods, seasonality, and trend." Other research, such as (Fricker et al., 2008b), has simulated background data using an additive combination of terms representing level, seasonal and day-of-week effects, and random noise.

In previous work we developed a multivariate simulation method which includes not only seasonal variation and day-of-week effects, but also allows for autocorrelation and cross-correlation structure in the data (Lotze et al., 2010). This allows for testing of multivariate methods which take advantage of the relationship between multiple

data streams, as well as creating data sets with realistic autocorrelation. This work, including R code and ten simulated data sets, is freely available at projectmimic.com.

(Siddiqi et al., 2007) developed a novel simulation method based on linear dynamical systems, also known as Kalman filters. They model the observed series as a linear transformation from a series of latent variables, find a stable linear transformation for those latent variables, and use this transformation to recreate similar data and to extend it into the future. They modify standard Kalman filter methods, incrementally adding constraints to create a system whose linear transformation remains stable (with eigenvalues less than 1). Most recently, (Maciejewski et al., 2009) developed a method which uses locally weighted regression (loess) to establish the total number of patients on each day (modeled by day-of-week effects, within-year components, long-term trend, and noise) and a multinomial model to determine symptoms. In addition, they add location, gender, and age to each simulated case, which allows testing of methods which take advantage of this additional information.

1.4.8. Outbreak Modeling

Being able to model outbreaks is important both for performing better algorithm comparison as well as for creating better detection algorithms. By creating more realistic models for the effect of an outbreak over time, one can inject more accurate simulated outbreaks; by incorporating these models into the detection algorithm, one could potentially have a more sensitive detection method.

There are few good recorded examples of actual outbreaks, aside from the yearly influenza outbreak (this is perhaps a likely reason for the increased attention to influenza detection in recent years). The classic example of an unexpected outbreak, studied retrospectively, comes from the limited data available from an anthrax outbreak in Russia (Meselson et al., 1994). Most research in this direction looks at modeling anthrax outbreaks (Brookmeyer et al., 2003, Brookmeyer et al., 2005), sometimes looking at its incubation period (Wilkening, 2008) or impact on grocery sales (Goldenberg et al., 2002b).

Other modeling research takes a more general approach. One direction is generating geographically based outbreaks and modeling spatial transmission (Watkins et al., 2007). Another very relevant approach is modeling the impact of actual disease occurrence and transmission on emergency department visits (Brillman et al., 2005). Similarly (Zhang et al., 2008) suggested a multivariate outbreak simulation method which derives multivariate aggregate data from simulated spatiotemporal cases, then estimating probabilities of seeking care along various indicators. A third approach is to combine outbreak and non-outbreak periods into a single model (Held et al., 2005). In determining the impact and evidence of a disease, the modeling of each step is important and can provide improved understanding of disease occurrence and detection effectiveness.

1.4.9. Spatial Detection Methods

In addition to methods which concentrate on the time series themselves, there has also been research into the use of geographic information. When the information

available is not simply a total count, but individual records from each patient (including geographical information such as home and/or work zip codes), one can attempt to determine not only whether or not there is an outbreak, but also where that outbreak might be within the monitored area. This can also allow more effective monitoring; by assuming that outbreak cases will have some geographic commonality, one can reduce false alerts from cases with no geographic consistency. This methodology seems very promising; however, while we will briefly review the literature in this area, this dissertation focuses on statistical detection using temporal pre-diagnostic data.

Kulldorff's scan statistic (Kulldorff, 1997, Kulldorff, 2001) is the basis for a majority of the research on spatio-temporal disease outbreak detection. Since its original proposal, it has been commented on (Lawson, 2001) and extended in various ways, such as multi-level spatial cluster analysis (Wallenstein & Naus, 2004, Que & Tsui, 2008), elliptical patterns (Kulldorff et al., 2003), irregularly shaped clusters (Duczmal et al., 2006), use on ordinal data (Jung et al., 2006), and use under a Bayesian framework (Neill et al., 2005, Neill et al., 2007). The spatial scan statistic has been applied to a number of biosurveillance problems including not only ICD-9 codes (Lazarus et al., 2002) but also West Nile detection via dead bird cluster analysis (Mostashari et al., 2003).

1.4.10. Other Biosurveillance-related Research

Other research has focused on other aspects of the biosurveillance process. For example, being able to automatically match patient names against possible near-

matches in a database (Bilenko et al., 2003, Jaro, 1995, Monge & Elkan, 1996), or automatically classifying hand-entered chief complaints into an ICD-9 category (Ivanov et al., 2003). Other areas of research include preserving privacy of health-related information while still being able to monitor it for epidemiological research. The general problem of confidentiality of data is dealt with by (Boyens, 2004, Dobra & Fienberg, 2001, Dobra et al., 2003, Domingo-Ferrer, 2002, Duncan et al., 2001), with applications more directly related to biosurveillance in (Fienberg, 2001). Much research is also in progress to improve physical disease detectors, either for hospital diagnosis or the creation of a city-wide array of aerosol detectors (Casman, 2004). While these last areas of research are clearly useful and contribute significantly to the success or failure of an actual biosurveillance implementation, they are beyond the scope of this dissertation.

1.5. Contributions of this Dissertation

The main contribution of this dissertation is threefold:

(1) to bring together many of the disparate approaches to biosurveillance,

(2) to introduce new improvements to algorithm development and evaluation, and

(3) to create a unified *statistical* framework.

In Chapter 2 we tie together forecasting and detection into a unified theoretical framework. In Chapter 3, improved forecasting methods are proposed and evaluated. Improved detection methods are proposed and evaluated in Chapter 4. Finally, in Chapter 5, we propose a method for evaluating how well simulated data captures the traits of the modeled series and also propose an improved visualization for daily

detection probability, thereby providing an improved framework for algorithm evaluation.

Chapter 2 : Forecast Accuracy and Detection Performance

2.1. Theoretical Framework

2.1.1. Problem Description

While many methods have been proposed for detecting disease outbreaks from pre-diagnostic data, their performance is usually not well understood. There is no theoretical framework for understanding why one method outperforms another, or why it works well on one type of data but not another. In this chapter, we describe a framework for providing this understanding, and show that it can effectively predict actual performance. The work in this chapter is based on previously published work in (Lotze & Shmueli, 2008b).

We begin to create such a framework by describing each detection algorithm as a combination of two components: a forecasting component coupled with a monitoring/detection component. In the forecasting component, the purpose is to provide an accurate forecast of the normal underlying health series behavior, which would be observed if there is no outbreak. The detection stage then takes the deviations from the forecast (or residuals), and applies a detection method, such as a Shewhart or CuSum statistic, in order to determine if the day is significantly higher than predicted. If the statistic is large enough, the system generates an alert for a health practitioner.

This decomposition into forecasting and monitoring describes nearly all biosurveillance detection methods described in Chapter 1. It clearly describes the standard methods of applying a detector directly to the data, where the forecaster is simply taken to be a constant. It also describes any method which pre-processes the data in an additive way. For example, the 7-day difference preprocessor can be seen as a forecasting method which takes as its forecast the value from 7 days ago. The detection algorithm is then applied to the residuals from this process, which are identical to the results of applying a 7-day differencing normalization to the baseline data. In several cases this sequence of "forecast, then monitor" is not done explicitly. Instead, a control chart is altered and then applied to the raw data. Even in such cases, the algorithm can be represented as a combination of forecasting and control chart monitoring. For example, in EARS or BioSense (programs initiated by the Centers for Disease Control and Prevention, see Section 1.2.2), a control chart is applied to the raw data, but a "sliding window" of recent data is used to set the control limits (as suggested in (Hutwagner et al., 2003)). This combination is equivalent to using a moving-average to forecast the next point and then applying a simple Shewhart control chart to the forecast errors. ESSENCE (a Department of Defense monitoring system, see Section 1.2.3) uses regression to forecast the next day's value, and then explicitly monitors the residuals in a control chart (described in (Lombardo et al., 2004)).

By examining the relationship between forecast accuracy and detection performance, we can put biosurveillance detection methods in a more general framework for

evaluation. Quantifying the effect of forecast precision on detection performance allows one to measure the benefits of improved forecasting and to determine when it is worth improving a forecast method's precision at a cost of robustness or simplicity. The effect of forecasting precision on Detection Rate is therefore applicable to biosurveillance, since it is important to know how much benefit improved forecasting will provide. Forecast methods are characterized by several important properties other than precision, such as robustness to non-normality, to outliers, or to outbreaks in the training data, as well as generating uncorrelated residuals (as discussed later in Section 2.5.2). When faced with a new forecast method which is more precise but is worse in, for example, robustness, the improvement must be quantified to understand the practical tradeoff.

Moreover, when residuals violate assumptions, e.g., when they are not independent or identically distributed, there are cases when *a better overall forecaster will actually have worse detection performance* for some kinds of outbreaks. By examining the effect of the residual properties on the detection performance, we can delineate these circumstances and better understand how to avoid them. If we can generate general rules about the effect of forecast precision on detection effectiveness, it will allow us to rank methods based on their actual forecast effectiveness, independent of the outbreak type or monitoring method. More importantly, quantifying this effect allows us to determine *how much* more effective the better forecast method will be, specific to the type of monitoring being applied and the type and size of the outbreak to be

detected. In addition, by examining properties of the residuals, we can identify those cases where a better forecast method will *not* necessarily produce better detection.

Although central to many applications, the effect of forecast precision on detection performance has not been directly studied. Monitoring and forecasting have been discussed as being similar in purpose and approach (by (Atienza et al., 1997)). The two also have been used together for the opposite purpose; rather than using forecasting to improve control chart detection, control charts have been used to identify issues in the forecast method, starting with (Van Dobben De Bruyn, 1967). In this chapter, we examine the quantitative effect of forecasting improvement on control chart detection, both in the standard case of independent identically distributed normal residuals as well as under various violations of assumptions which occur in practice.

2.1.2. Problem Formalization

We first consider a series with no outbreak signals; we call such a series the underlying background or baseline series, denoted as u_t (t=1,2,…). It is this underlying background that a forecast method is attempting to forecast. The predictions from the forecast method are f_t; if we examine the forecast errors, $e_t = u_t - f_t$, we can estimate the Mean Squared Error (MSE) or Root Mean Squared Error (RMSE) and bias of those errors. This will be useful in evaluating detection effectiveness.

Since we do not actually know a priori whether or not the data contain an outbreak, we denote the actual values in the series as y_t. When there is no outbreak signal, $y_t = u_t$. Let o_t be the outbreak signal at time t. In general, $y_t = u_t + o_t$, which assumes an additive number of cases due to the outbreak signal. For most days, $o_t = 0$, whereas $o_t > 0$ only on days where there is an outbreak. This reflects the epidemiological model commonly used in biosurveillance. If a multiplicative outbreak effect is assumed ($y_t = u_t o_t$, where $o_t > 1$ only on days where there is an outbreak), we can use a log transform and model $\log(y_t)$ instead of y_t, thereby converting to an additive outbreak form.

Since we do not know if an outbreak is present in a given series, we will refer to the difference $r_t = y_t - f_t$ simply as a residual, rather than a pure forecast error. In the absence of an outbreak signal, r_t will be a pure forecast error and the residuals will have variance equal to the forecast method's MSE (assuming unbiased forecasts). However, in the presence of an outbreak signal, r_t will contain an additional term; since the forecast method is forecasting only the underlying background, we will not call this a forecast error. The residual can thus be separated into two components, $r_t = (u_t - f_t) + o_t$. The first component is the forecast error ($e_t = u_t - f_t$) and the second is the outbreak signal (o_t).

An illustration of these components can be seen in Figure 2-1. It shows the original series and forecasts in the left panel; the residuals obtained from subtracting the forecasts from the original series are shown in the right panel.

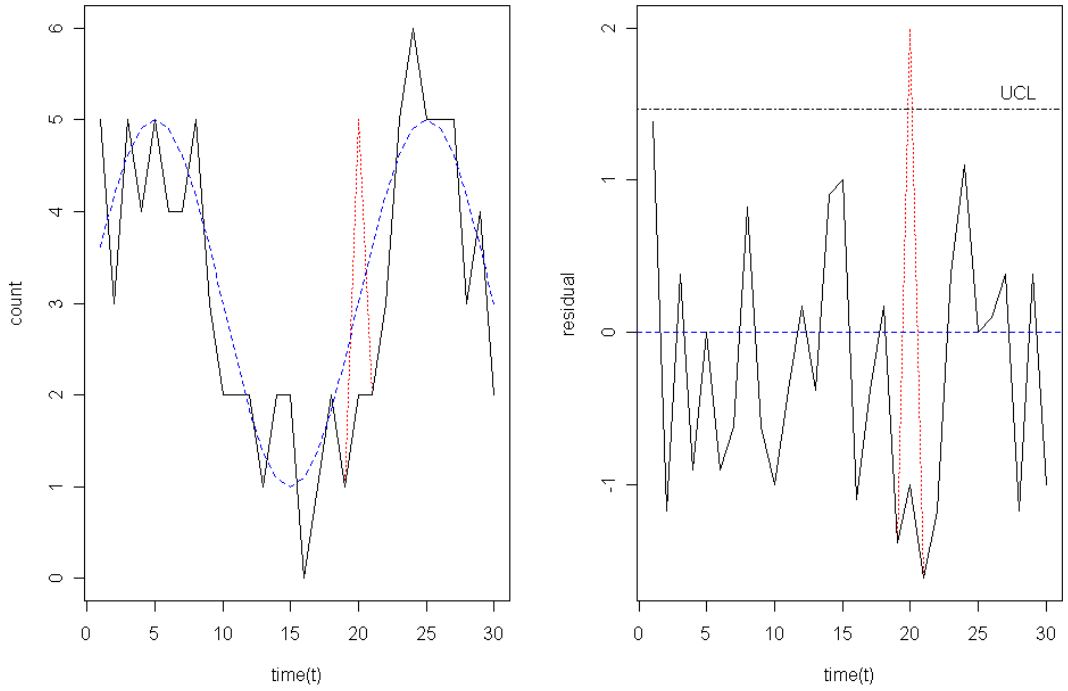

Figure 2-1: Illustration of Forecasting and Detection
The left panel shows an original series (black solid line, u_t) and its forecasts (blue dashed line, f_t). The right panel shows the residuals from subtracting forecasts from the series, in a one-sided Shewhart control chart. The red dotted line is the addition of an outbreak signal ($o_t + u_t$).

2.2. *The Idealized Case*

2.2.1. Gaussian iid Residuals with Mean 0

In our analysis, we first assume that the forecast method generates forecast errors with a given MSE. Initially, we assume that these errors are independent, normally distributed, with mean 0 and constant variance. We later relax these assumptions and re-evaluate performance.

We now consider an additive outbreak signal that is injected into the monitored series. This outbreak signal is considered to be independent of the background or residuals. Thus, we are in the realm of standard control charts: we are seeking a

change in the process mean, given a series of independent identically distributed (iid) normal observations. Let us first consider a single-day 'spike' outbreak signal.

Note that when converting a time series to a series of residuals, if the residuals have 0 mean, then the residuals' variance is equal to the forecast method's MSE.

2.2.2. Detection

First, consider a one-sided Shewhart chart being applied to residuals that are iid, $e_t \sim N(0, \sigma^2)$. Setting the upper control limit at UCL means that a false alert will occur with ATFS

$$ATFS = \frac{1}{1 - \Phi(UCL/\sigma)}. \qquad \text{(Eq. 2-1)}$$

In the simplest case, the outbreak signal is of constant size, $o_t = \eta$. In this case, the algorithm will detect if $e_t/\sigma + \eta/\sigma > UCL/\sigma$. By using the same transformation as above, the control chart will correctly alert on the day of the outbreak if $Z > UCL/\sigma - \eta/\sigma, Z \sim N(0,1)$, which translates into a Detection Probability equal to

$$\text{Detection Rate} = 1 - \Phi(UCL/\sigma - \eta/\sigma). \qquad \text{(Eq. 2-2)}$$

Note that we obtain Equation 2-1 by setting $\eta = 0$.

Now consider two forecast methods, f_1 and f_2 with RMSEs equal to σ_1 and σ_2, respectively, and where $\sigma_1 < \sigma_2$ (i.e., forecast method f_1 provides more precise forecasts). If detectors on each of f_1 and f_2 are set to have the same false alert rate ($ATFS_1 = ATFS_2$) we can write $UCL_1/\sigma_1 = UCL_2/\sigma_2$. Denote this level as *a*:

$UCL_1/\sigma_1 = UCL_2/\sigma_2 = a$. Since $\sigma_1 < \sigma_2$, then clearly $UCL_1 > UCL_2$. Thus the corresponding probabilities of detection will be $TA_1 = 1 - \Phi(a - \eta/\sigma_1)$ and $TA_2 = 1 - \Phi(a - \eta/\sigma_2)$. Because $\sigma_1 < \sigma_2$ and Φ is monotonically increasing, we get $1 - \Phi(a - \eta/\sigma_1) > 1 - \Phi(a - \eta/\sigma_2)$, and thus $TA_1 > TA_2$. Therefore the more precise forecast method (f_1) will also provide a higher Detection Rate.

The effects are shown in Figure 2-2, where the Detection Rate of five forecast methods are compared, all normalized to have the same ATFS. We see that as the forecasting becomes more precise (i.e., the RMSE decreases), the Detection Rate increases. While this relationship is monotonic (a lower RMSE always results in improved detection), the amount of improvement depends on the *size* of the outbreak signal (η). Since $UCL = \sigma\Phi^{-1}(1 - 1/ATFS)$ (see Equation 2-1), the improvement in Detection Rate from using f_1 over f_2 can be expressed as

$$\Phi(\Phi^{-1}(1 - 1/ATFS) - \eta/\sigma_2) - \Phi(\Phi^{-1}(1 - 1/ATFS) - \eta/\sigma_1). \quad \text{(Eq. 2-3)}$$

Due to the nature of the normal cumulative distribution function Φ, this quantity must be computed numerically.

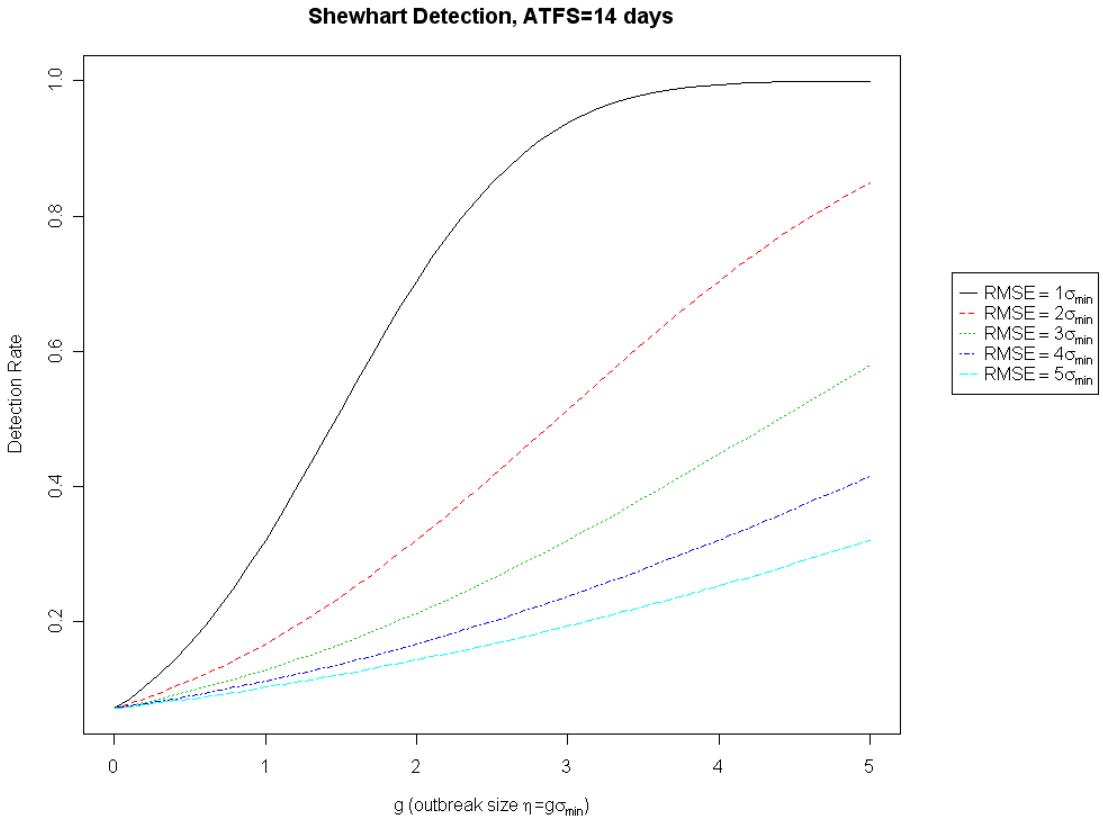

Figure 2-2: RMSE Effect on Shewhart Detection
Comparison of Shewhart chart performance for forecast methods with different RMSEs, as a function of outbreak size ($g = \eta/\sigma_{min}$, where σ_{min} is the RMSE of the best forecast method).

We compute similar probabilities for EWMA charts in Section 2.6.1.

2.2.3. Timeliness

When outbreak signals last more than one day, there are more chances to detect them. This allows consideration not only of the probability of detection, but also the distribution of *when* the outbreak is detected.

We first consider a fixed step increase of size η that starts at time *i* and continues indefinitely ($o_i = \eta, \forall i > t$). Such an outbreak signal could be the result of an environmental contamination (biological or chemical) resulting in a constant increase

in the number of illness cases. Since any control chart method will eventually alert, we focus on timeliness over true alert probabilities. In control chart terminology, this is usually referred to as the Average Run Length (ARL), which is the expected number of days until an alert is generated.

For the Shewhart chart, each day is essentially a Bernoulli trial in terms of detection, with probability of success $p = 1 - \Phi(UCL/\sigma - \eta/\sigma)$. Thus, the number of days until detection is a geometric random variable with expected value $ARL = (1-p)/p$. (If the alerting day is considered to be included, then $ARL = (1-p)/p + 1 = 1/p$.)

The relationships between outbreak size and expected delay (i.e., the number of days until detection), for forecast methods of varying precision, can be seen in Figure 2-3. Results for EWMA and CuSum charts are in Section 2.6.1 and Section 2.6.2, respectively. Note that the *quantity* of the performance difference varies significantly based on the outbreak size and the amount of forecast improvement; the amount of improvement is crucial in determining the practical benefits from using an improved forecast method.

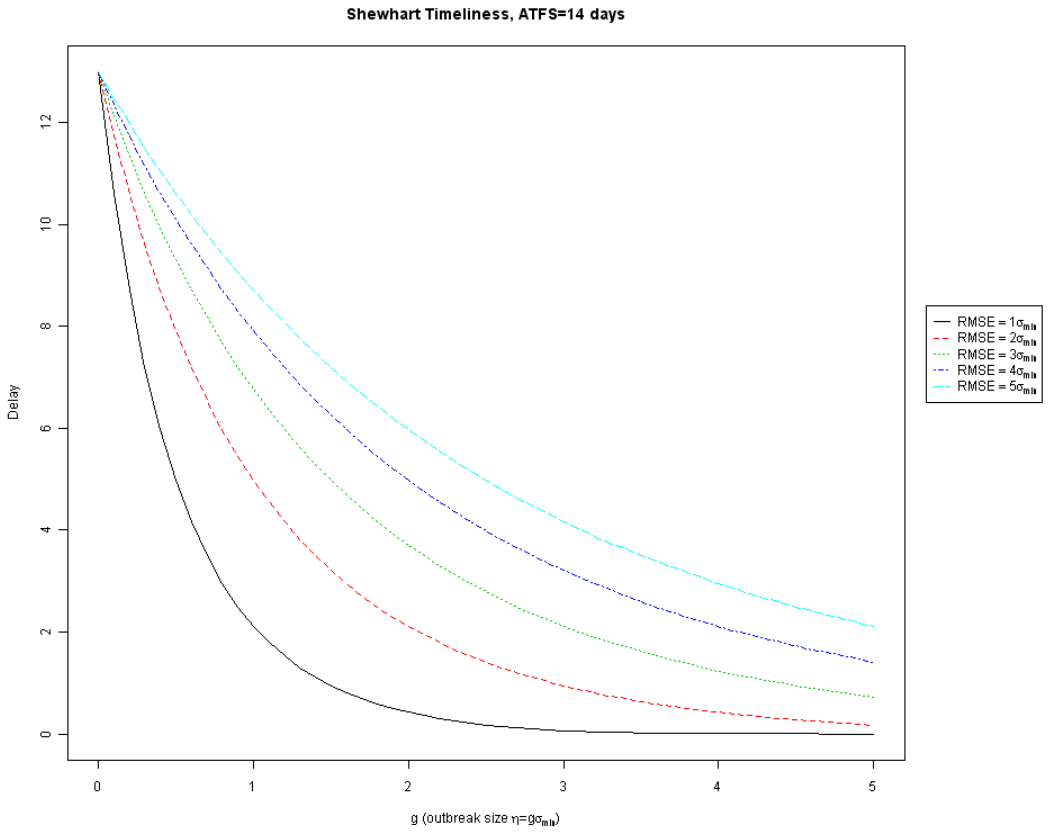

Figure 2-3: RMSE Effect on Shewhart Timeliness
Comparison of Shewhart chart timeliness for forecast methods with different RMSEs, as a function of outbreak size ($g = \eta/\sigma_{min}$, where σ_{min} is the RMSE of the best forecast method).

We caution that in practice the expected value (ATFOS) may not be the most useful metric, since it will incorporate alerts that were generated many days after the outbreak signal first appeared in the data. In other words, it averages over the entire distribution of possible delays. If a detection must occur within the first k days of an outbreak signal in order to be useful to the user, then more effective metrics of model performance and comparison are the probability of alert *within the first k days* and the *conditional expected timeliness*, given that an alert occurred within the first k days. This same issue comes up when recognizing the finite duration of outbreaks; if an outbreak only lasts k days, then a detection must certainly occur within k days to be

useful. In essence, one must make sure to examine detection probability as the probability of practically useful detection, and timeliness as the expectation of delay, conditional on a practically useful detection.

An important condition of our results regarding improved forecasting leading to improved detection is that the forecast method does not include the outbreak in the background data and thereby forecast the combination of background plus outbreak (a problem described in (Burkom et al., 2007)). This can be achieved in practice by using a 'guardband window' which means that forecasts are generated for more than one day ahead. Forecasting farther into the future generally results in reduced precision, which in turn leads to deteriorated detection probabilities and timeliness. It is, in fact, precisely when considering tradeoffs of this kind that one must quantify the loss from decreased forecast precision.

2.3. Unknown Residual Distribution

2.3.1. Bounds for Residuals with Unknown Distribution

If the residuals have mean 0 and variance σ^2, but their distribution is unknown, then we can use a Chebyshev inequality to bound the detection probability. We will also here sometimes use the terms False Alert (FA), where FA=1/ATFS and True Alert (TA), where TA=Detection Probability. We know that at least $1 - 1/k^2$ of the values are within k standard deviations from the mean. This means that we can guarantee a false alert rate FA by setting:

$$UCL = \frac{1}{\sqrt{FA}}\sigma \qquad \text{(Eq. 2-4)}$$

Note that already this is conservative: if the distribution is symmetric, then half of the values outside k standard deviations will be low, and so the false alarm rate will only be FA/2. In practice, the UCL should be set by empirical estimation of the distribution (using past residuals to determine a UCL that obtains a specified FA).

The condition for alerting is $r_t = e_t + o_t > UCL$. Given a distribution for $r_t = e_t + o_t$, we can compute TA by integrating the distribution of r_t over the area above the control limit, and the actual FA by integrating the distribution of e_t above the control limit. However, when the distribution is unknown, the Chebyshev bound on detection means that η must be strictly larger than the UCL to guarantee detection, as the bound gives no guarantee on the probability of occurrence above the mean. Thus, the probability of detection is bounded by

$$1 - \frac{1}{\left(\frac{\eta - UCL}{\sigma}\right)^2} = 1 - \frac{\sigma^2}{(\eta - UCL)^2} \qquad \text{(Eq. 2-5)}$$

Figure 2-4 shows the relationship between reduced RMSE and improved detection, even with very conservative bounds.

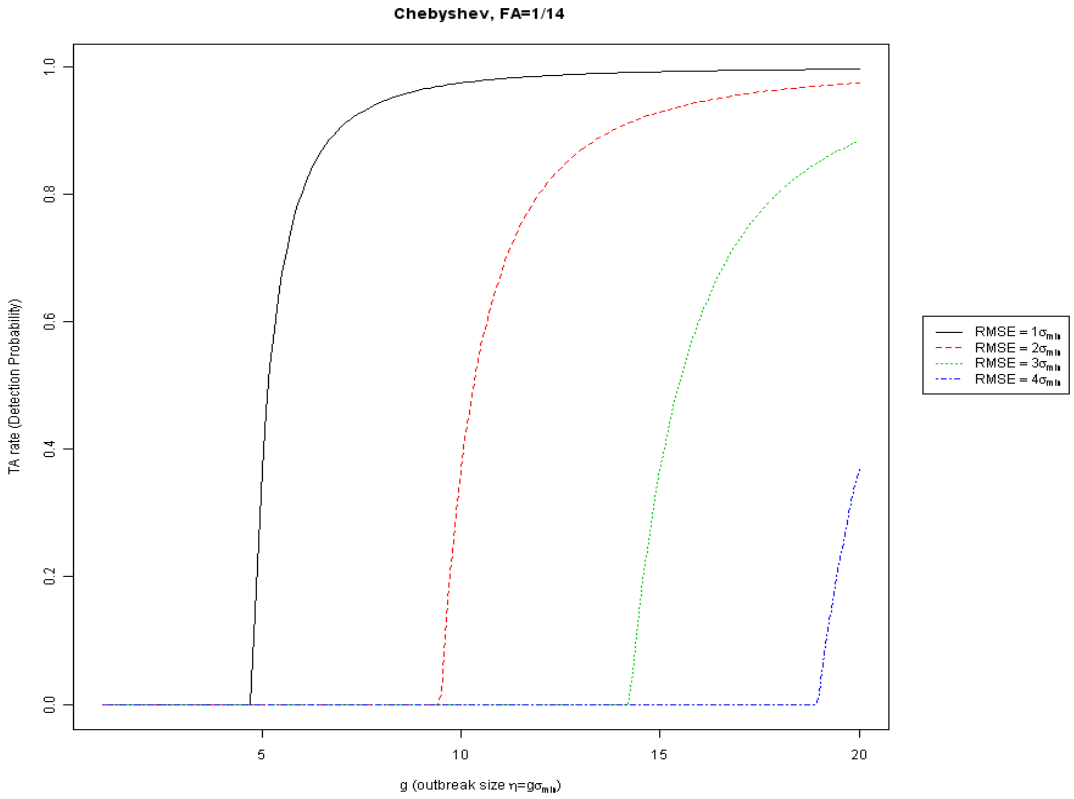

Figure 2-4: Chebyshev Bounds for Detection
Comparison of lower bounds on Shewhart chart performance for forecast methods with different RMSEs, where the residual distribution is unknown.

Obviously, this bound is not a very good one, nor very tight in reality (although using the one-sided Chebyshev inequality could be used to tighten this bound, for most distributions, the detection probability will approach 1 much faster than seen here). However, it does again show that regardless of the distribution of residuals, improved forecasting leads to improved detection.

2.4. Extension to Stochastic Outbreaks

2.4.1. Importance of Stochastic Outbreak Analysis

In a real-life disease surveillance scenario, an outbreak will not be of fixed size and location. Instead, it will have a variable impact, infecting more or fewer people

depending on uncontrollable factors like traffic, social interactions, or work intensity. In addition, a variable number of people will report their symptoms, purchase over-the-counter remedies, or contact a health advisor. It is thus more appropriate to think of an outbreak as a stochastic realization of a random process, and recognize that the outbreak signal will likewise be stochastic.

Research in comparing biosurveillance techniques generally does not take this stochastic outbreak signal into account (exceptions include work by Burkom simulating stochastic lognormal outbreaks (Burkom & Murphy, 2007a) and WARE's CityBN simulator (Wong et al., 2005)), but as we show below, the stochastic nature of an outbreak can have a significant impact on the performance of different methods.

2.4.2. Gaussian Stochastic Outbreak

We examine and quantify the impact of a stochastic outbreak signal on detection performance. Departing from normal control chart assumptions, we assume that the outbreak is not of fixed size, but is instead stochastic, e.g., $o_t \sim N(\eta, \nu^2)$. In this case, a Shewhart chart has probability of detection equal to

$$DetectionRate = 1 - \Phi\left((UCL - \eta)/\sqrt{\sigma^2 + \nu^2}\right). \qquad \text{(Eq. 2-6)}$$

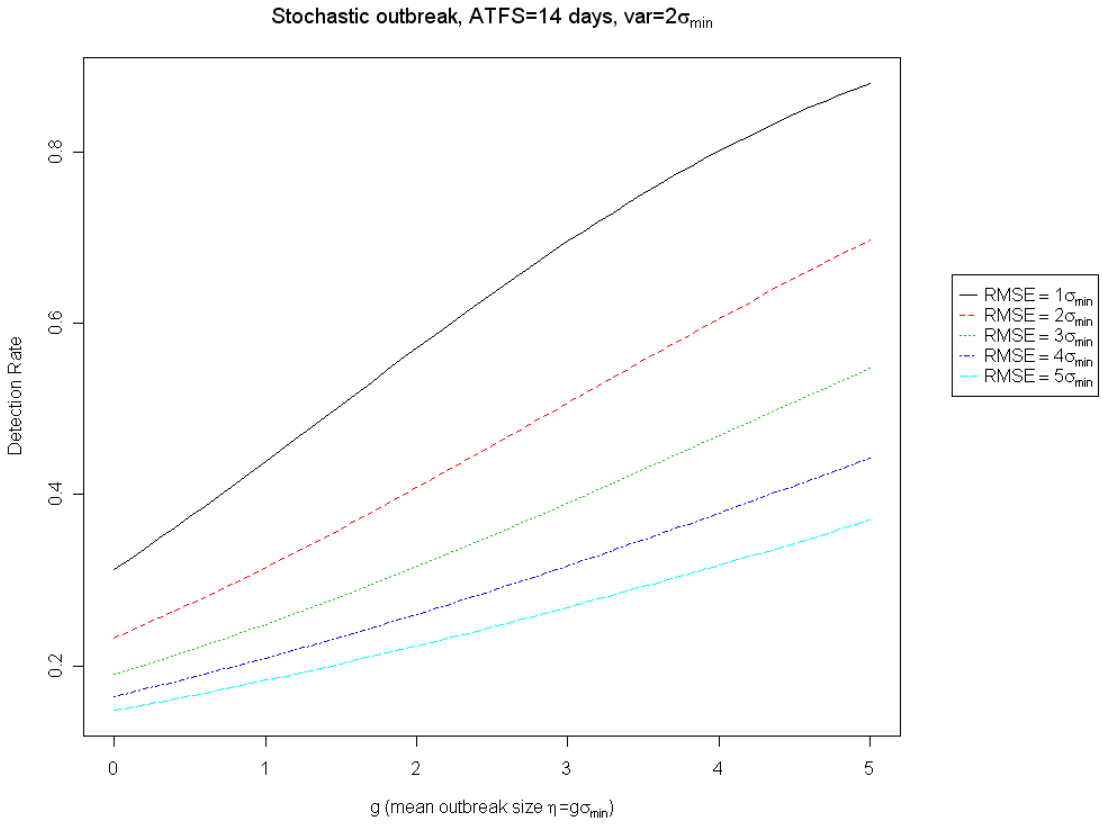

Figure 2-5: Stochastic Outbreak Performance
Detection rate if the outbreak is a stochastic Gaussian spike outbreak with mean $\eta = g\sigma_{min}$ and variance $2\sigma_{min}$. Each line indicates a forecaster with different accuracy.

Figure 2-5 shows the relationship between expected outbreak size (η) and Detection Rate for a stochastic outbreak signal, applying a Shewhart control chart to five forecast methods with varying RMSEs. Compared to the fixed-size spike (as seen in Figure 2-2), the increased variance in the outbreak signal reduces the Detection Rate for larger spikes, but increases it for smaller ones; this can be clearly seen in Figure 2-6, which shows directly the change in Detection Rate if the outbreak is stochastic rather than fixed.

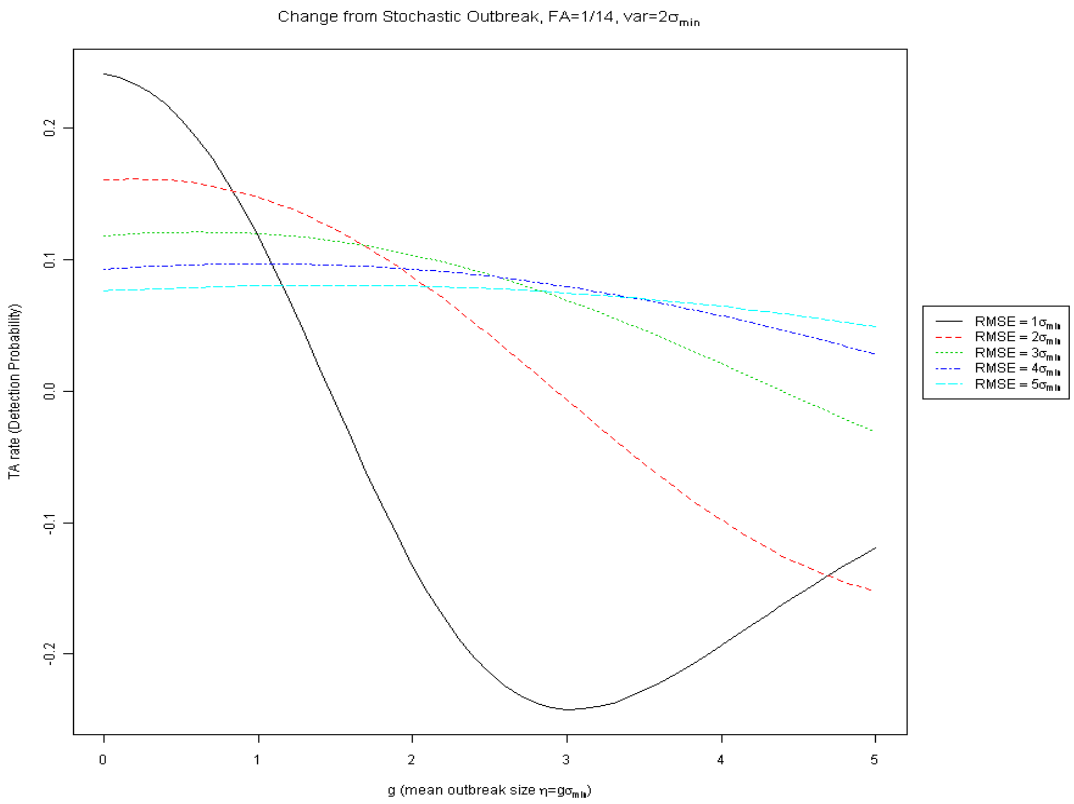

Figure 2-6: Performance Change due to Stochastic Outbreak
Change in Detection Rate if the outbreak is stochastic rather than fixed, for a Gaussian spike outbreak with variance $2\sigma_{min}$. Each line indicates the effect on a forecaster with different accuracy.

The effect is proportional to the amount of outbreak-size variance, ν^2. In comparing two methods, this distortion can drastically affect the relative performance of the two forecast methods. A large advantage of one forecast method over another under constant variance may be almost *trivial* under a different outbreak-size variance.

2.5. *Extensions to Day-of-week Seasonal Variance and Autocorrelation*

2.5.1. Day-of-week Seasonal Variance

When the forecast precision is non-constant, even if the forecast method produces unbiased forecasts, the theoretical analysis in Section 2.2 does not hold. This can occur, for example, when the series of daily counts follows a Poisson distribution

with different λ parameters for each day of the week. In this case, even if the mean value is correctly forecasted, the variance of the residual will depend on the day's λ parameter. A similar effect can occur when an additive forecast method is applied to a series with multiplicative background behavior. Although a preliminary log transformation of the series may be a reasonable solution, such a transformation will also have a significant impact on the outbreak signal.

Seasonal variance can also be induced by deseasonalizing methods which normalize values by multiplication. An example is deseasonalizing a series from a day-of-week effect using the ratio-to-moving-average method (as described in (Lotze et al., 2008)). But if such methods are used appropriately, they may help reduce seasonal variance by normalizing the variance of residuals across seasons. However, here too there is the danger that a transformation that affects the variance of the residuals will also impact the size of the outbreak signal.

If there is periodic variance in the residual series with period k, we can represent the variance as a set of variances, $\{\sigma_1^2 ... \sigma_k^2\}$. Then the overall variance of the series (assuming that the mean residual=0 for each season) is $\sum_{i=1}^{k} (1/k)\sigma_i^2$. If the seasonal pattern is such that some days have equal variance, we can represent this as $\sum_{i=1}^{k} \alpha_i \sigma_i^2$, where α_i is the proportion of days with variance σ_i^2. Given this mixture model for seasonal variance, we can compute the probability of detection. For a step outbreak signal using a Shewhart control chart, we can compute separate probabilities

of detection by season; thus, the probability of detection for an outbreak signal of size η is

$$DetectionRate = \sum_{i=1}^{k} \alpha_i P(detection|\eta, \sigma_i). \qquad \text{(Eq. 2-7)}$$

Using Equation 2-2, this quantity is equal to $\sum_{i=1}^{k} \alpha_i (1 - \Phi((UCL/\sigma_i) - (\eta/\sigma_i)))$, where the UCL is derived from the overall variance of the series.

As an example relevant to biosurveillance, consider the case of a forecaster which has high variance on weekdays, but lower variance on weekends, as in Figure 2-7.

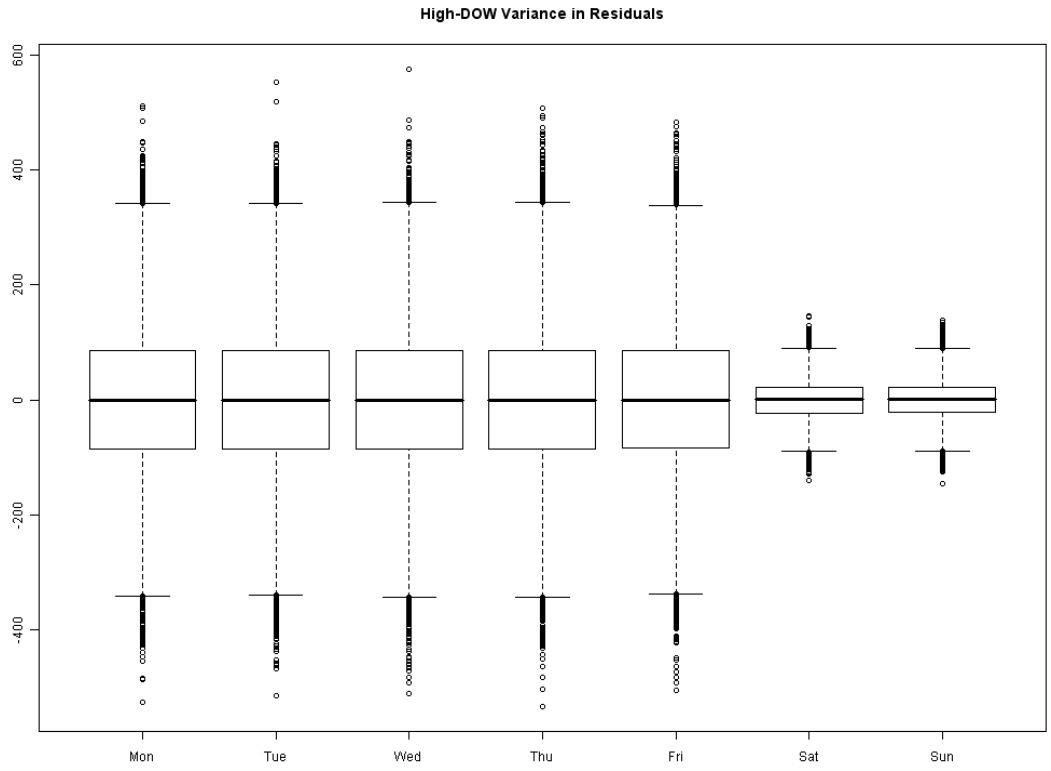

Figure 2-7: Box-and-whiskers Plot of Seasonal Variance
A typical example of seasonal variance in biosurveillance residuals. The box-and-whiskers plots show the median, 25% and 75% percentiles as a box, the range of the (non-outlier) remaining data as whiskers, and further outliers as individual points. It can be seen that the residual variance is much lower on weekends than on weekdays (due largely to lower counts on weekends).

For this scenario, the detection probability is:

$$(5/7)(1 - \Phi(UCL/\sigma_{weekday} - \eta/\sigma_{weekday})) +$$
$$(2/7)(1 - \Phi(UCL/\sigma_{weekend} - \eta/\sigma_{weekend})).$$

If the overall variance is kept constant at 100, but the difference between weekend and weekday variance is increased, the performance becomes more markedly different from the constant variance case. We can see this difference in performance in Figure 2-8; as weekday and weekend variances become more distinct, Detection Rates deteriorate for small outbreak sizes, but actually *improve* for some intermediate

outbreak sizes. At these intermediate outbreak sizes, the increased probability of detection when the outbreak occurs on low-variance weekends outweighs the decrease in performance on higher-variance weekdays. As the overall variance is increased, this "kink" pattern of deviation from the constant variance case is increased.

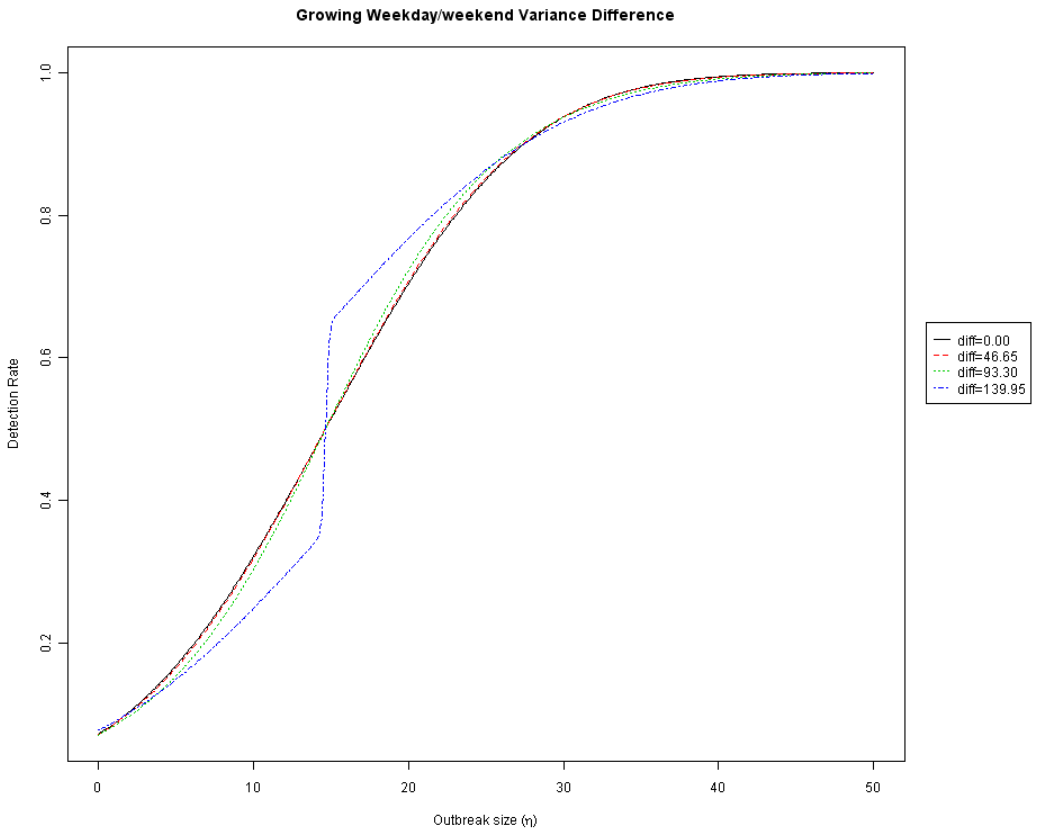

Figure 2-8: Seasonal Variance Effect on Shewhart Detection
Shewhart chart performance for forecast methods with identical overall variance ($\sigma^2 = 100$), but different residual seasonal variances (diff=difference between weekday and weekend residual variance).

If variance is strongly differentiated by season, an improved RMSE will not always give better detection performance, depending on the size of the outbreak. For some outbreak sizes, a forecast method with a larger overall RMSE but low weekend RMSE can outperform a forecast method with a smaller overall RMSE. When there

is significant seasonal variance, the performance can be evaluated more accurately using Equation 2-7 and estimates for the different seasonal variances. This suggests that improved monitoring can be achieved by using different UCLs and/or different forecast methods for each season.

2.5.2. Autocorrelation

Autocorrelation in a series of residuals means that the residuals on consecutive days are linearly correlated. Autocorrelated residuals indicate that the forecast method did not capture part of the dependence structure in the raw data (such as a seasonal component). In biosurveillance data, the most pronounced autocorrelation in series of residuals is that of lag 1 (the correlation between r_t and r_{t-1}) and it is typically positive. This can arise in practice because the yearly seasonality has not been completely accounted for (and so still has a residual effect on neighboring residuals) or because the data arise from an ARMA process which has not been correctly modeled by the forecast method. When we refer to autocorrelation hereafter, we are referring to positive autocorrelation.

When data are autocorrelated, the series will have increased variance due to the autocorrelation. In the case of an autoregressive model of order 1 (AR(1)), given by

$$y_t = \phi y_{t-1} + \epsilon_t, \epsilon_t \sim N(0, \sigma_z^2), \qquad \text{(Eq. 2-8)}$$

the resulting variance is $\sigma_z^2/(1-\phi^2)$ (Maragah & Woodall, 1992). The effect of autocorrelation on detection performance has been examined in the control chart literature. Several papers that look at Shewhart, CuSum, and EWMA charts applied to

autocorrelated series indicate that autocorrelation leads to a greater number of false alarms, due to the greater variance in the series (Maragah & Woodall, 1992, Woodall & Faltin, 1993, Padgett et al., 1992, Noorossana & Vagjefi, 2005) . However, for Shewhart charts, if the control chart limits are adjusted to account for the variance of the actual autocorrelated series (rather than the variance which would exist without any autocorrelation), then the overall probability of detection will remain the same for a spike outbreak. We do caution that while this is true unconditionally, the probabilities of detection, conditional on the value for the previous day, are not identical for each day. The probability of alert will be larger on days following large values, and smaller on days following small values. As we discuss below, this implies that methods taking this conditional probability into account should provide improved detection performance.

Although the performance of a Shewhart chart is unaffected by autocorrelation on single-day spike outbreaks, there will be a longer average delay in detection when considering a multi-day outbreak signal, both for Shewhart and other control charts. When the outbreak begins on a day with a small residual, which is too low to trigger an alert (even after the outbreak signal addition), the subsequent residuals will likely also be too low for the outbreak to be detected. Thus, average delay will increase for higher values of autocorrelation. These effects are shown in Section 2.7.1.

To determine whether or not a residual series contains autocorrelation, an autocorrelation (ACF) plot may be used (with α-level bounds $(z_{1-\alpha/2})/\sqrt{N}$). When

autocorrelation is present, as mentioned above, the conditional probability of detection varies by day; this implies that one might use an ARMA-type or other model as an additional forecasting step on the residuals from the original forecast method (such models are described in (Box & Luceno, 1997, Montgomery & Mastrangelo, 1991)). However, note that in the case of multi-day outbreaks, such models will incorporate the outbreak signal into the forecasting, and thus the assumption of independence of outbreak and forecast error will be violated. The results of such incorporation on the performance of detection algorithms is discussed in (Hong & Hardin, 2005). The decrease in performance from incorporating an outbreak must be measured against the gain achieved by reducing the autocorrelation, as mentioned at the end of Section 2.2.3; it is precisely these kinds of tradeoffs for which this theoretical quantification is useful.

2.6. Extension to CuSum and EWMA Charts

2.6.1. EWMA Chart

We can measure the effect of improved forecasting on EWMA chart detection, as in Equation 2-3 for Shewhart charts, by noting that $EWMA_t$ is a normal random variable, with mean 0 and variance as in (Montgomery, 2001):

$$\sigma^2_{EWMA_t} = \sigma^2 \left(\frac{\lambda}{2-\lambda} \right) \left(1 - (1-\lambda)^{2t}\right), \qquad \text{(Eq. 2-9)}$$

where t is the number of time points since the EWMA was started. After an initial startup period, the variance converges to $\sigma^2 \left(\lambda/(2-\lambda)\right)$. The one-sided EWMA chart has been shown to have very similar performance to the EWMA approximated

by this steady-state normal distribution (Shu et al., 2007). By an argument similar to the Shewhart case (in Section 2.2), we can show that the improvement in detection probability from using forecaster f_1 over f_2 can be expressed as

$$\Phi\left(\Phi^{-1}(1-\frac{1}{ATFS})-\frac{\lambda\eta}{\sigma_2\sqrt{\frac{\lambda}{2-\lambda}}}\right)-\Phi\left(\Phi^{-1}(1-\frac{1}{ATFS})-\frac{\lambda\eta}{\sigma_1\sqrt{\frac{\lambda}{2-\lambda}}}\right)$$

Note that if λ=1, this simplifies to Equation 2-3.

Figure 2-9 shows the relationship between outbreak size (η) and Detection Rate for EWMA detectors when applied to five different forecast methods, each with a different RMSE.

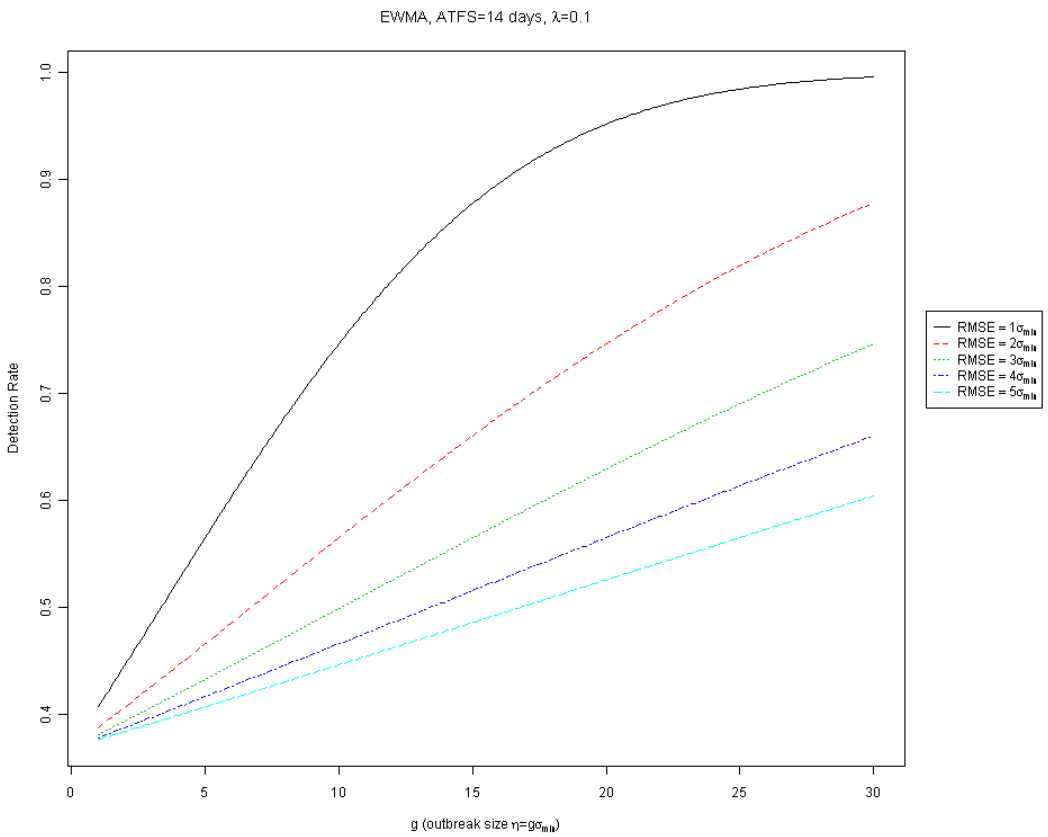

Figure 2-9: RMSE Effect on EWMA Detection

Comparison of EWMA chart performance for forecast methods with different RMSEs, as a function of outbreak size ($g = \eta/\sigma_{min}$, where σ_{min} is the RMSE of the best forecast method.).

Comparing Figure 2-2 and Figure 2-9 shows that an EWMA chart has a lower chance of detecting a spike outbreak compared to a Shewhart chart with the same ATFS, when both are applied to residuals with the same RMSE. The reason is that by giving the maximum weight to the most recent observation, the Shewhart chart is more tuned to detect spike outbreak signals. A much larger spike is necessary to achieve the same Detection Rate with an EWMA chart. However, we also note that as a weighted sum of observations, the EWMA statistic is more robust to deviations from normality, and so may be more effective when the residual distribution is further from normal.

We can also examine the impact of detection on the timeliness of the EWMA chart. For the EWMA chart, the ARL is computed numerically; we use the method described in (Crowder, 1987), numerically integrating the Fredholm equation using Gaussian quadrature.

The relationships between outbreak size and expected delay (i.e., the number of days until detection), for forecast methods of varying precision, can be seen in Figure 2-10.

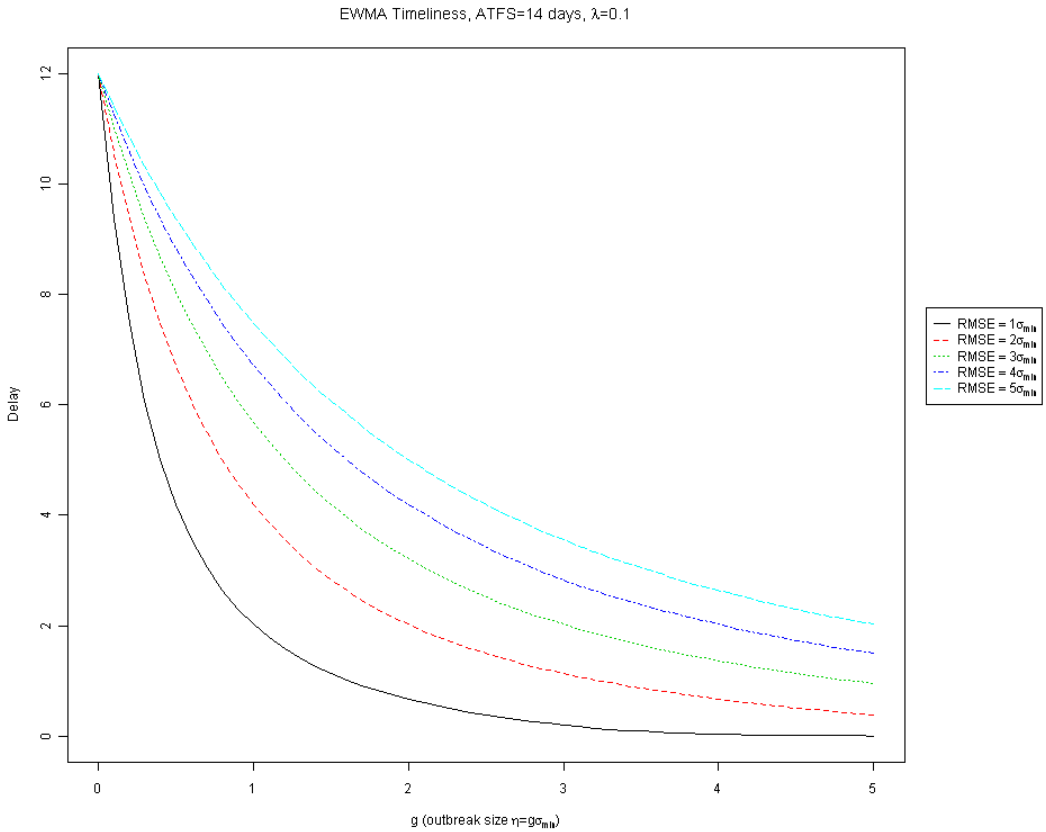

Figure 2-10: RMSE Effect on EWMA Timeliness
Comparison of EWMA chart timeliness for forecast methods with different RMSEs, as a function of outbreak size ($g = \eta/\sigma_{min}$, where σ_{min} is the RMSE of the best forecast method.) As with Shewhart charts, more precise forecasts result in faster detection.

2.6.2. CuSum Chart

In a CuSum chart, unlike the Shewhart chart, the monitoring statistics on different days are no longer independent, and therefore the number of days until an alert is no longer follows the geometric distribution. However, the ATFS can still be accurately determined using numerical methods or approximations. One such approximation is found in (Siegmund, 1985), which approximates the ATFS by

$$ATFS \approx 2(e^{-2(UCL/\sigma+1.166)} + UCL/\sigma + .166). \qquad \text{(Eq. 2-10)}$$

This same approximation can provide the ATFOS:

$$ATFOS \approx \frac{e^{-2\Delta b} + 2\Delta b - 1}{2\Delta^2}, \qquad \text{(Eq. 2-11)}$$

where $\Delta = \eta/\sigma - 1/2$ and $b = UCL/\sigma + 1.166$.

The relationships between outbreak size and expected delay (i.e., the number of days until detection), for forecast methods of varying precision, can be seen in Figure 2-11. As with Shewhart and EWMA, more precise forecasts result in faster detection.

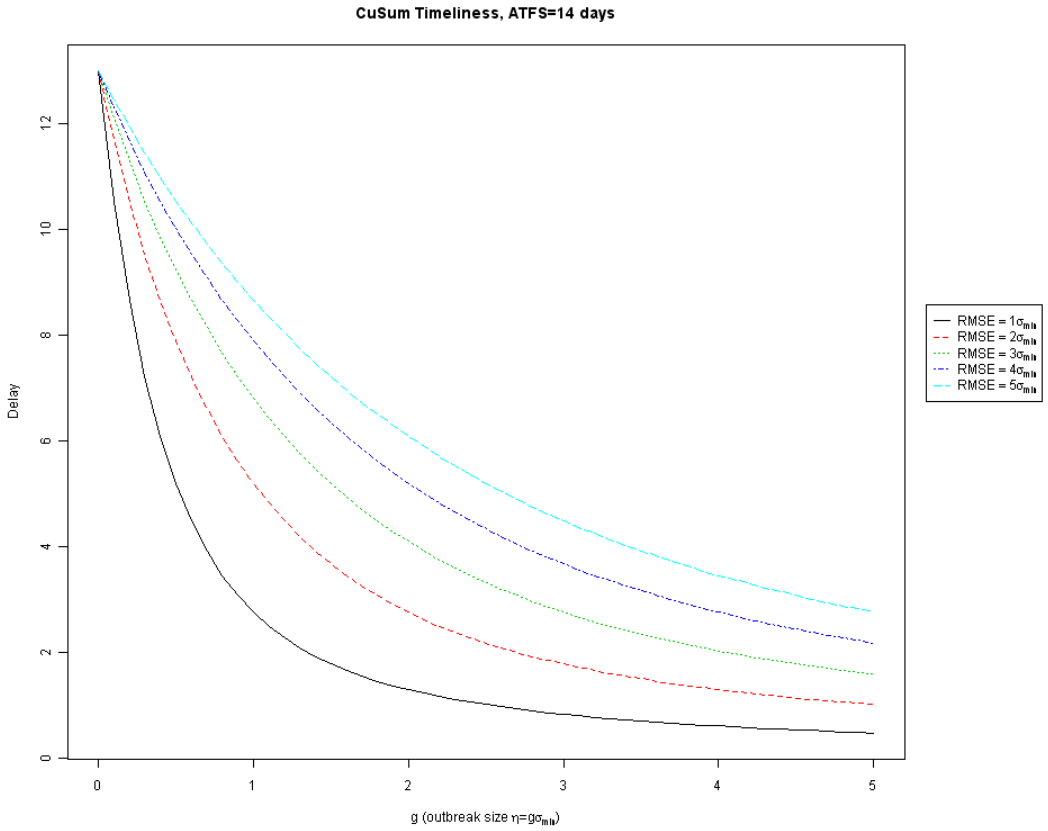

Figure 2-11: RMSE Effect on CuSum Timeliness
Comparison of CuSum chart timeliness for forecast methods with different RMSEs, as a function of outbreak size ($g = \eta/\sigma_{min}$, where σ_{min} is the RMSE of the best forecast method.)

2.6.3. Comparison of CuSum and Shewhart Charts

One surprising result, as seen in Figure 2-12, is that although the CuSum chart has improved timeliness over the Shewhart chart for small outbreak signals (as expected), the Shewhart chart quickly catches up and outperforms the CuSum as the outbreak size increases. In addition, this timeliness improvement appears to be bounded below, and to hold only for a certain range of outbreak sizes.

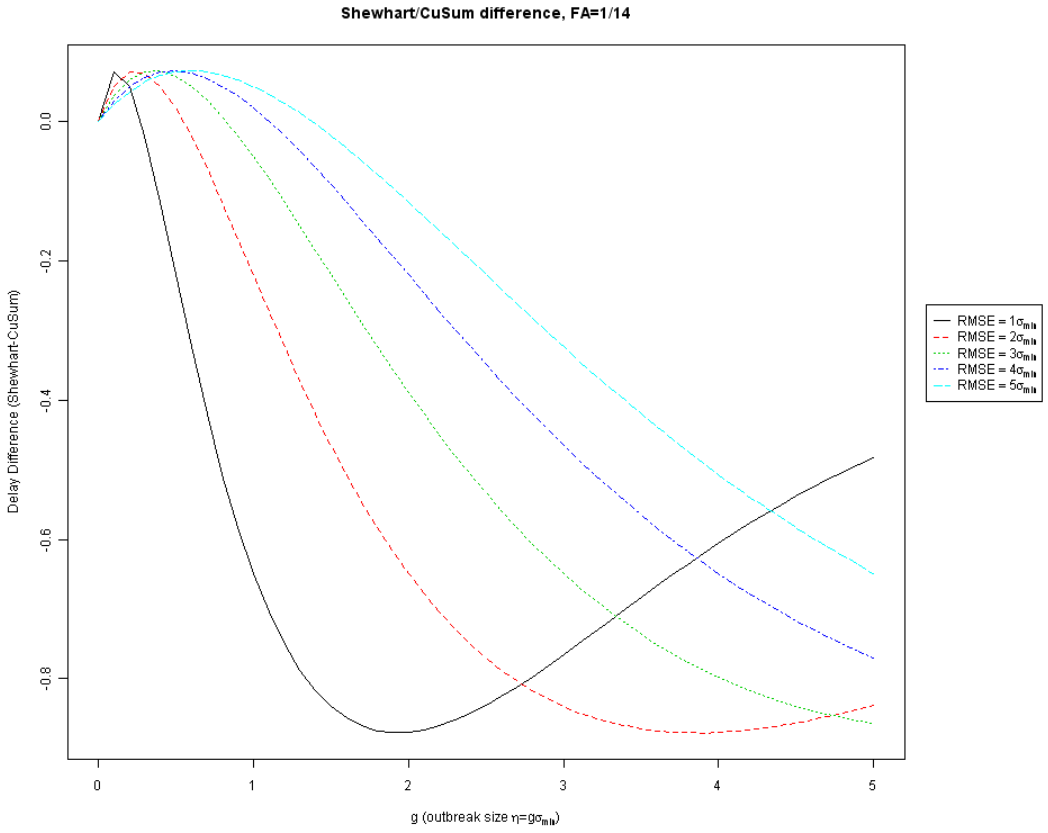

Figure 2-12: Timeliness Differences Between Shewhart and CuSum
Expected difference in delay resulting from using a Shewhart instead of a CuSum chart, on the same forecast residuals. When the value is negative, the Shewhart chart provides faster expected detection than CuSum.

We can see that at this false alert level, for step outbreaks, there is little reason to use a CuSum chart over a Shewhart chart. This appears to conflict with the result from (Fricker et al., 2008b), in which the CuSum is shown to be significantly more

powerful at detecting outbreaks than Shewhart. However, this will be resolved in Section 5.3.3, by examining the false alert levels using heatmaps derived from the quantitative detection analysis from this chapter.

2.7. Empirical Confirmation of Theoretical Results

2.7.1. Autocorrelation Simulations

To study the impact of autocorrelation on detection and timeliness performance, residuals were simulated using different levels of autocorrelation, but again maintaining the same overall variance. In the Shewhart charts using spike outbreaks, no significant deviation was seen from the theoretical performance, when the control limit was set according to the final resulting variance. Figure 2-13 shows that the detection performance is not affected by autocorrelation.

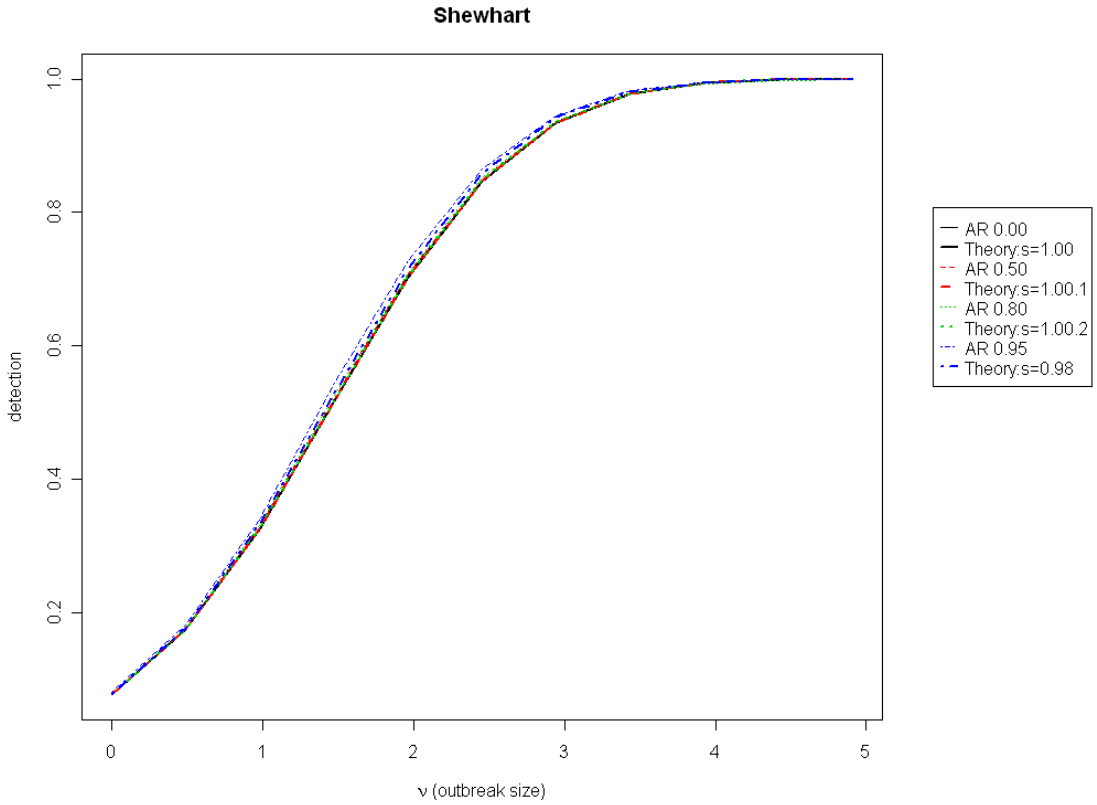

Figure 2-13: Autocorrelation Effect on Shewhart Detection
Shewhart chart performance for forecast methods with different residual autocorrelation levels (ACF) but identical overall variance $\sigma^2 = 1$).

Figure 2-14 shows a significant deterioration in timeliness for small outbreak sizes and high autocorrelation. This is in agreement with (Wheeler, 1991, Wheeler, 1992) regarding the relatively small impact of most autocorrelation levels on Shewhart chart performance.

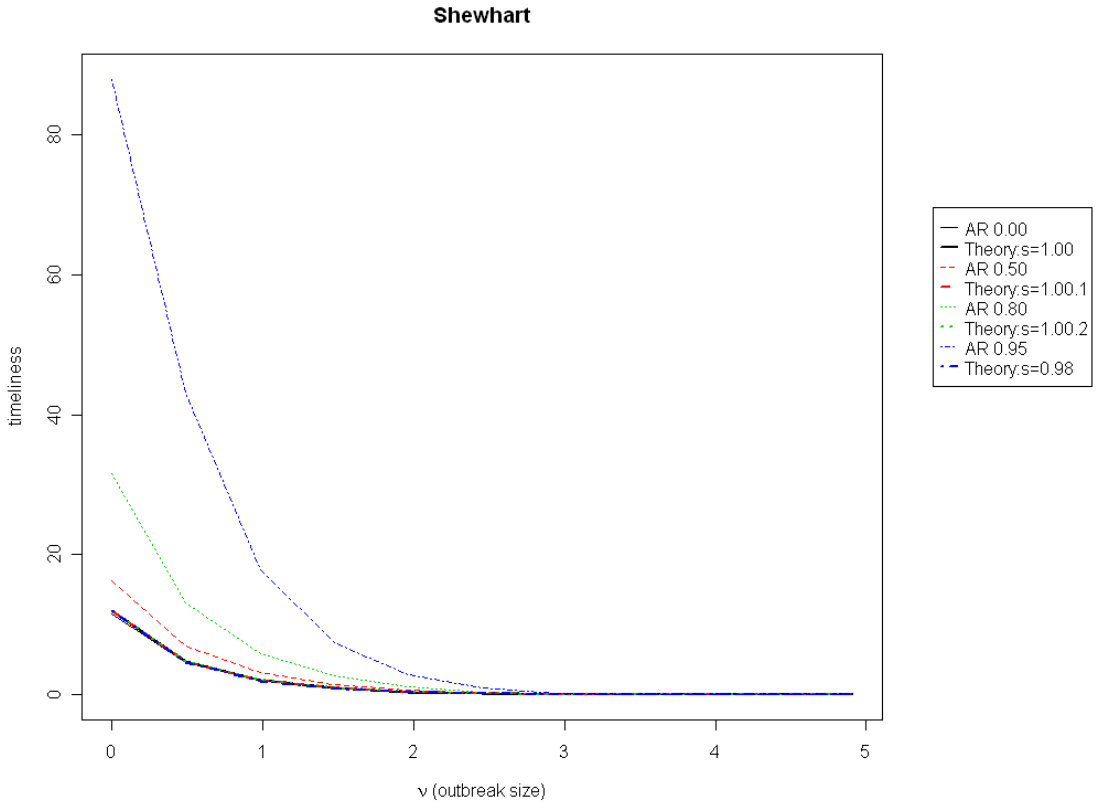

Figure 2-14: Autocorrelation Effect on Timeliness
Shewhart chart timeliness for forecast methods with different residual autocorrelation levels (ACF) but identical overall variance $\sigma^2 = 1$).

Results using CuSum charts on autocorrelated data are similar to those for Shewhart charts. Detection may be slightly affected for small spike outbreaks (as seen in Figure 2-15), and timeliness is more strongly affected than in Shewhart charts (as seen in Figure 2-16).

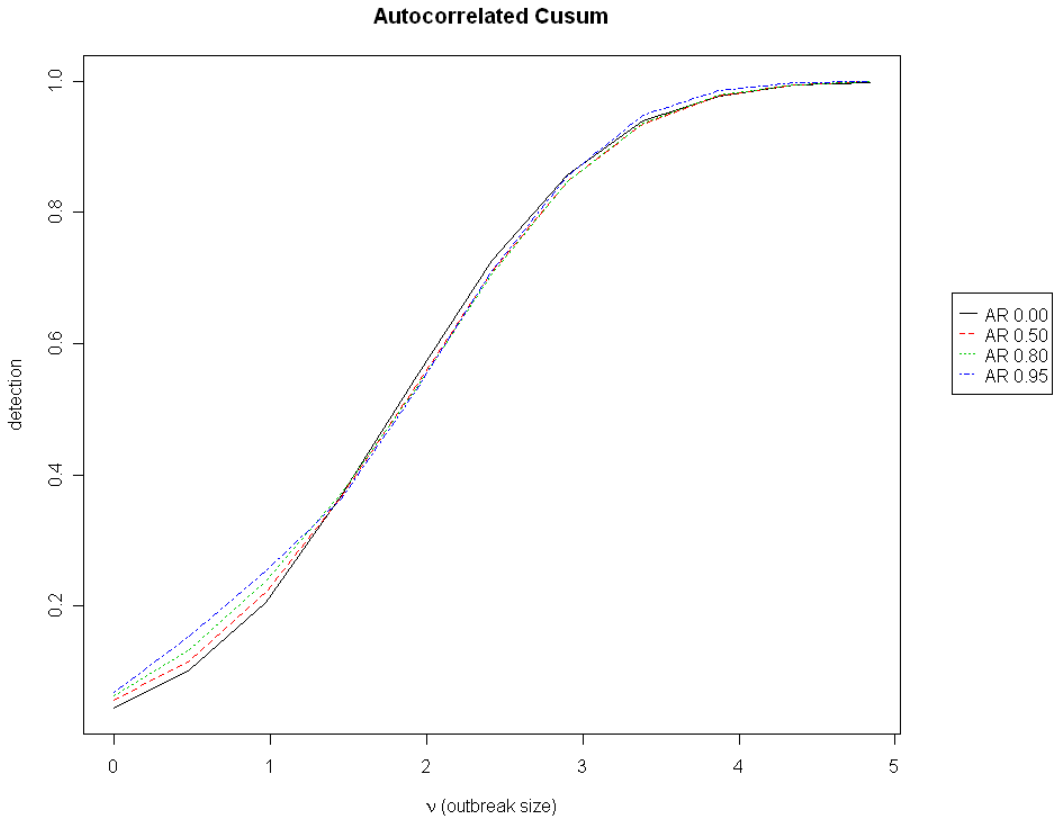

Figure 2-15: Autocorrelation Effect On CuSum Detection

Empirical Detection Rate of CuSum charts applied to residuals with the same overall variance, but different levels of autocorrelation. The x-axis shows the size of a spike outbreak signal, and the y-axis shows the probability of detection.

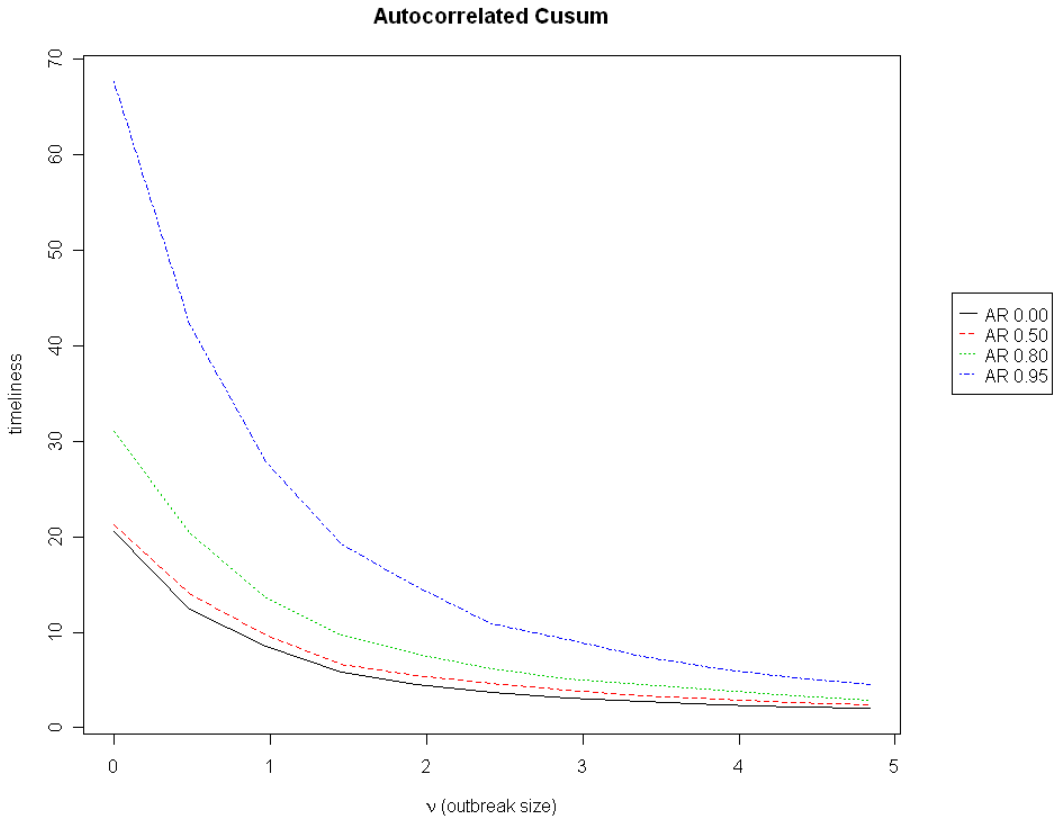

Figure 2-16: Autocorrelation Effect on CuSum Timeliness
Empirical timeliness of CuSum charts applied to residuals with the same overall variance, but different levels of autocorrelation. The x-axis shows the size of step outbreak signal, and the y-axis shows the probability of detection.

2.7.2. Application to Authentic Data

An authentic health data set is now used to determine the effectiveness of theory when estimating performance of currently-used forecast methods. These tests show the applicability of the theory to the evaluation of forecast methods on actual health data for detecting disease outbreaks. If the predicted performance and actual performance match well, then the theoretical analysis can be used to accurately estimate the detection performance of actual systems; thus, the forecast metrics can be a useful comparison metric, without requiring computationally intensive simulation studies.

To examine the forecast methods' effectiveness, authentic health series data are used, with a simulated outbreak signal inserted at various possible dates of outbreak. This methodology is now commonly used in biosurveillance to estimate the effectiveness of detection (Goldenberg et al., 2002a, Reis & Mandl, 2003, Stoto et al., 2006). The technique involves using an authentic health data set from a health provider, simulating a potential outbreak signal and inserting the simulated additional counts in the authentic data. Then, the detection algorithm is run to determine whether it alerts during the simulated outbreak, and if so, how quickly. By repeating this routine multiple times and inserting the simulated outbreak at multiple points, one can estimate how the detection algorithm would perform during an actual outbreak.

For this validation, we use data from the BioALIRT program conducted by the U.S. Defense Advanced Research Projects Agency (DARPA), described in Section 1.3.1. For this study, we use the daily count of respiratory symptoms from civilian physician office visits, all within a particular U.S. city. The first 1/3 of the data (233 days) was used for training, and the last 2/3 (467 days) for evaluation.

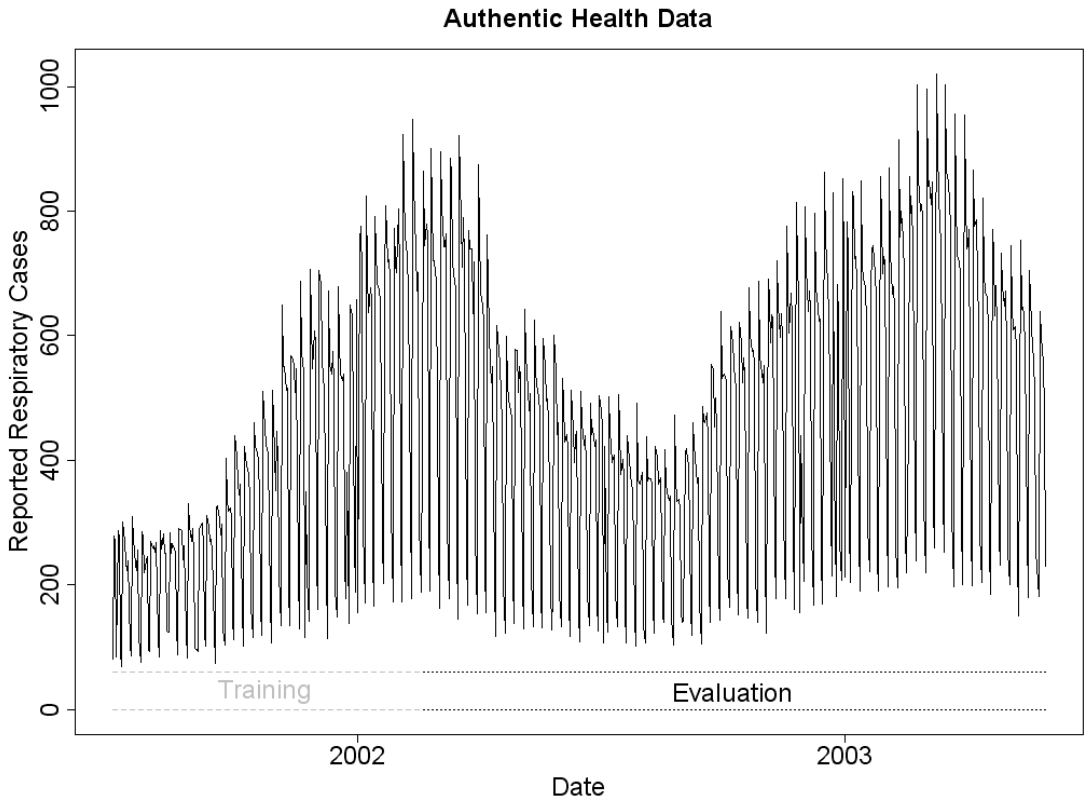

Figure 2-17: BioALIRT Civilian Respiratory Data
Original data series, split into sections for training and evaluation.

Simulated spike outbreak signals of various sizes (0-300 additional cases) were generated and inserted into every day in the evaluation set, creating 467 trial data sets for each outbreak signal size. For each outbreak size, the Detection Rate was calculated as the average over all 467 insertions. An illustration of the process can be seen in Figure 2-18.

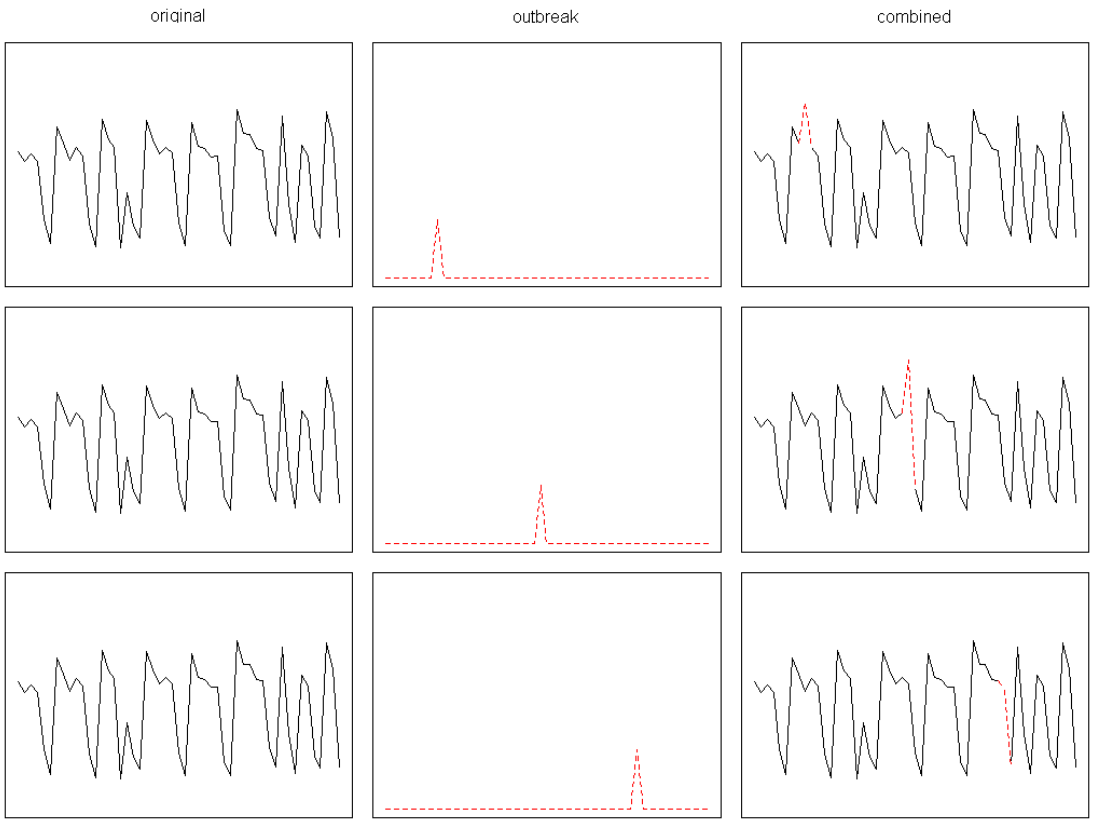

Figure 2-18: Outbreak Injection Example
Illustration of taking raw, authentic health data series and injecting a spike outbreak into three different days, resulting in three test data series. These evaluation series are then used as outbreak-labeled time series for estimating the method's Detection Rate. In our implementation, 467 such data series were created for each outbreak size.

Three forecast methods for forecasting next-day daily counts were compared: Holt-Winters exponential smoothing, 7-day differencing, and linear regression. For a more detailed description of these methods, see Section 3.2. For each method, the first 1/3 of the data (233 days) was used for training, and the last 2/3 (467 days) for evaluation. Note, however, that the 7-day differencing has no real "training" to speak of, and that both the Regression and Holt-Winters method incorporate *all* previous days when generating a forecast.

The RMSE for each forecast method was computed in the training data. This RMSE was used to generate a theoretical performance curve for each forecast method as described in Section 2.2. Actual performance was computed using the method described in Section 2.1, using the forecast method for prospective forecasting, subtracting the forecast to generate residuals, and applying a Shewhart control chart to those residuals.

Results can be seen in Figure 2-19, which compares the actual performance from a forecasting method's residuals to the performance which would be expected from the theoretical performance for residuals of the same overall RMSE. When a constant UCL was used, the actual performance was somewhat similar to that predicted by theory, but seemed to underdetect small outbreaks and overdetect midsized outbreaks. This result is similar to that seen under seasonal variance (see Section 2.5.1), which reflects the seasonal variance of the residuals (seen in Figure 2-20).

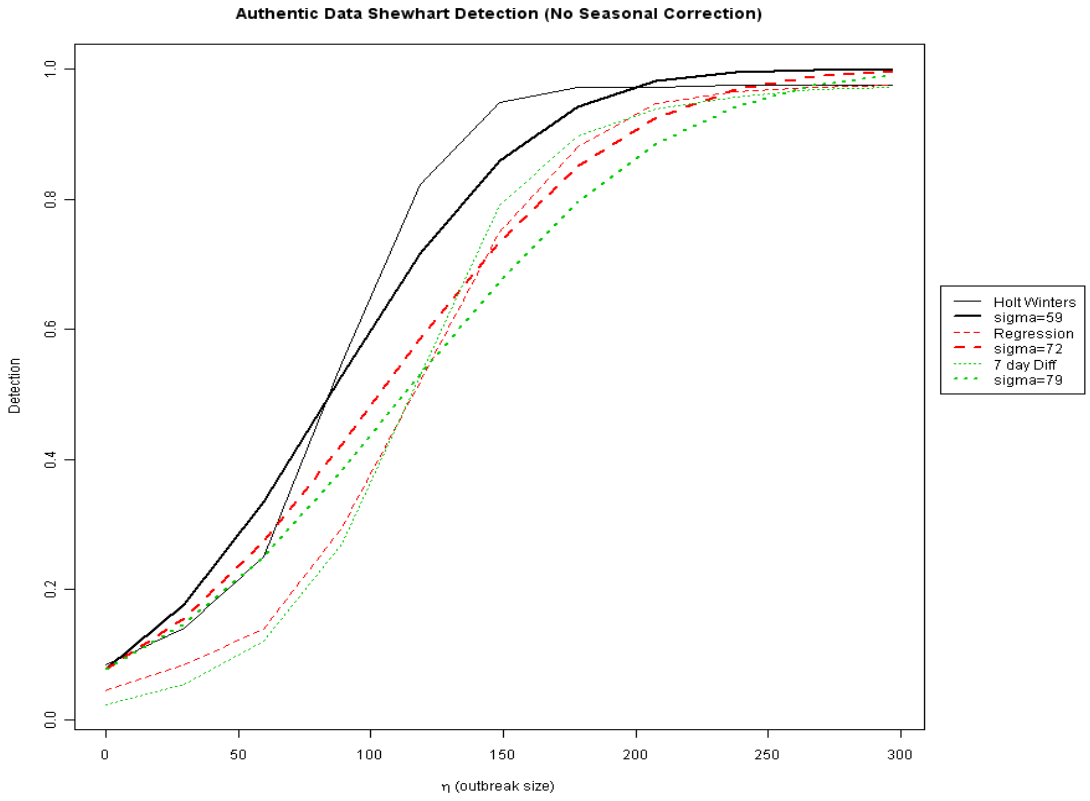

Figure 2-19: Empirical Shewhart Detection Performance

Actual (thin) and theoretical (thick) Shewhart chart performance for forecast methods with different RMSEs, assuming constant variance, as a function of outbreak size (η). Solid/black=Holt-Winters, Dashed/red=7-day Diff, Dotted/green=Regression. Each forecasting method has the σ for its residuals measured, and is matched with a plot of theoretical performance for residuals of the same σ.

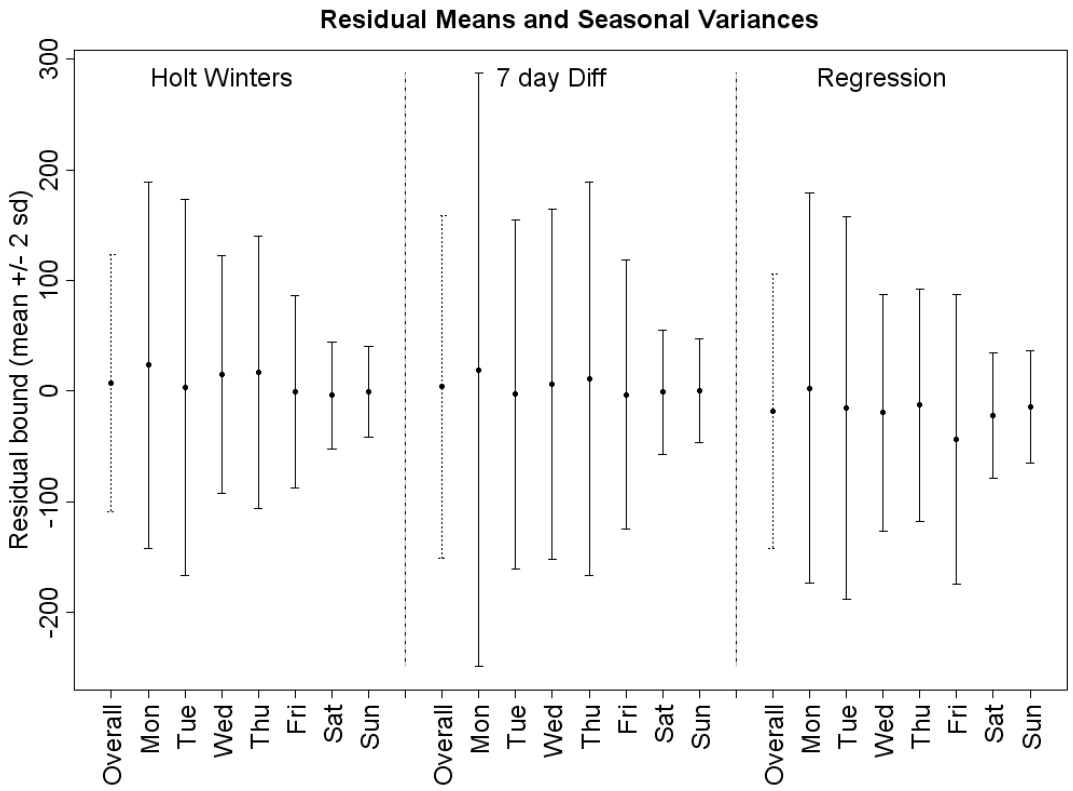

Figure 2-20: Residual Means and Seasonal Variance
Residual variance of the three forecast methods, and variance by day-of-week. Seasonal day-of-week variance affects detection performance, and can be accounted for by using the formulas in Section 2.5.1.

A further examination was done, with variance computed for each day-of-week and performance predicted using seasonal variance computations. The results are shown in Figure 2-21, where an improved fit is seen, especially for the Holt-Winters residuals, although there is still some difference on the larger outbreaks.

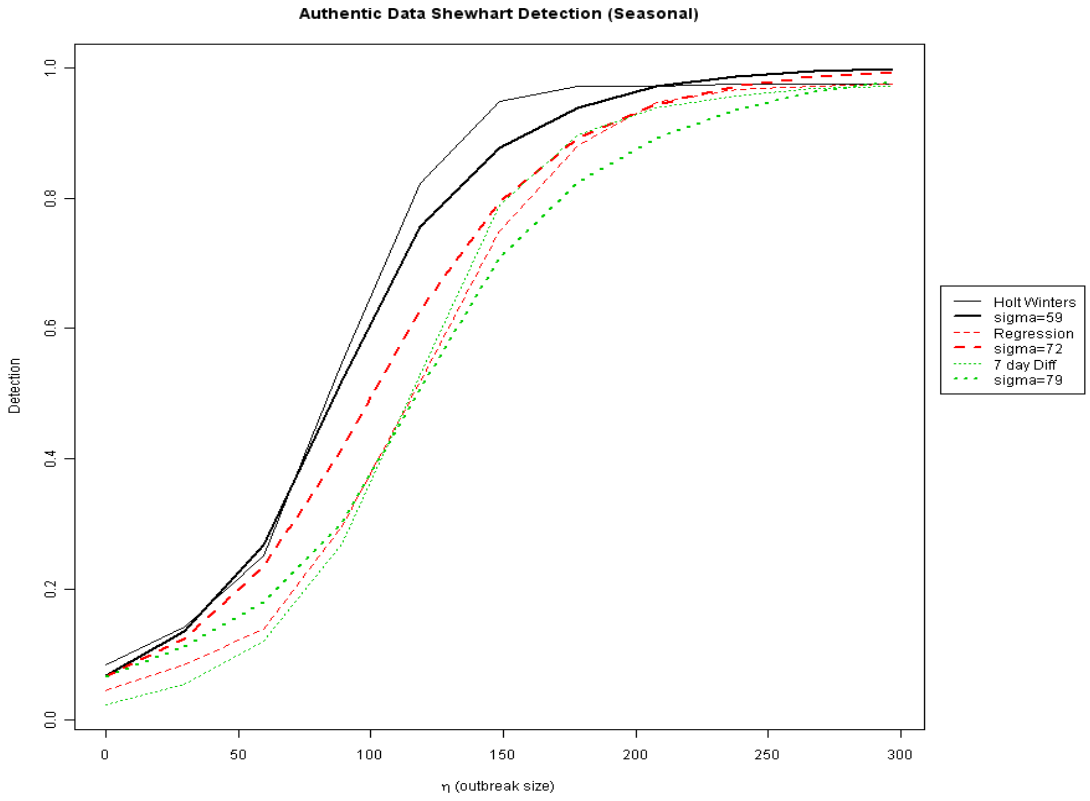

Figure 2-21: Empirical Shewhart Detection Performance With Seasonal Variance
Actual (thin) and theoretical (thick) Shewhart chart performance for forecast methods with different RMSEs, assuming day-of-week variance, as a function of outbreak size (η). Solid/black=Holt-Winters, Dashed/red=7-day Diff, Dotted/green=Regression. Each forecasting method has the σ for its residuals measured, and is matched with a plot of theoretical performance for residuals of the same σ, with the same day-of-week residual variance.

This improvement is quantified in Table 2-1, which presents, for each method, the average percentage error when using the constant variance assumption versus using the seasonal variance correction.

Method	Nonseasonal	Seasonal
Holt-Winters	4.10%	2.70%
7 day Diff	4.89%	4.41%
Regression	6.36%	4.07%

Table 2-1: Average Percentage Error With or Without Seasonal Correction
The average percentage error in predicted detection rate for each method using the theoretical framework, over outbreak sizes from 0 to 300. For each method, using the day-of-week seasonality adjustment results in a more accurate estimate of detection probability.

In order to compare timeliness, the experiment was repeated using step outbreaks instead of spike outbreaks. Step outbreaks have an additional count which begins on a certain day, and lasts indefinitely. Figure 2-22 compares the timeliness performance of real forecast methods to theoretical performance predicted by a 7-day seasonal variance model. The timeliness is worse for small outbreaks, particularly for the regression and 7-day differencing.

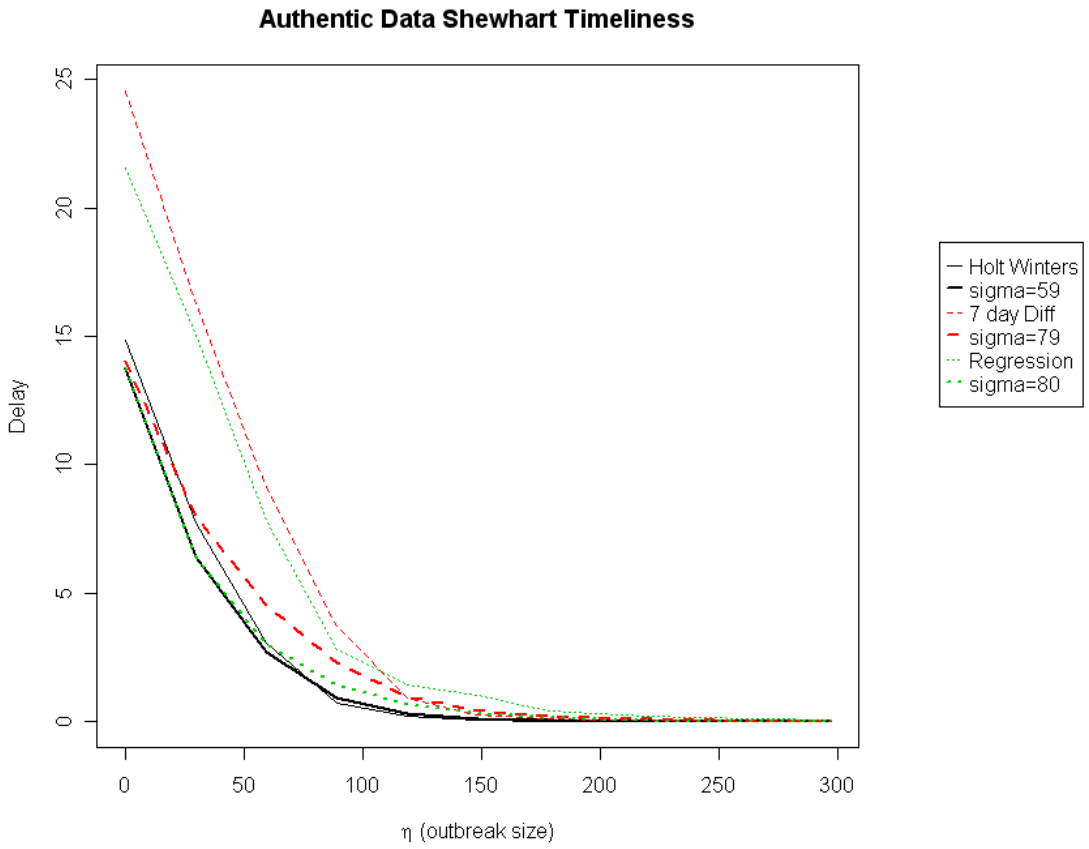

Figure 2-22: Empirical Shewhart Timeliness Comparison
Actual (thin) and theoretical (thick) Shewhart chart timeliness for forecast methods with different RMSEs, assuming constant variance, as a function of outbreak size (η). Solid/black=Holt-Winters, Dashed/red=7-day Diff, Dotted/green=Regression. Each forecasting method has the σ for its residuals measured, and is matched with a plot of theoretical performance for residuals of the same σ, with the same day-of-week residual variance.

The extra delay for regression and 7-day differencing seems to be due to autocorrelation: as seen in Figure 2-23, the regression and 7-day differencing residuals have larger autocorrelation than Holt-Winters. Alternatively, the overall differences may be due to the bias of the residuals (none has mean 0) or their non-normal distribution. In spite of these, we see that the forecast methods' performance ranking is related to their RMSE ranking, as expected.

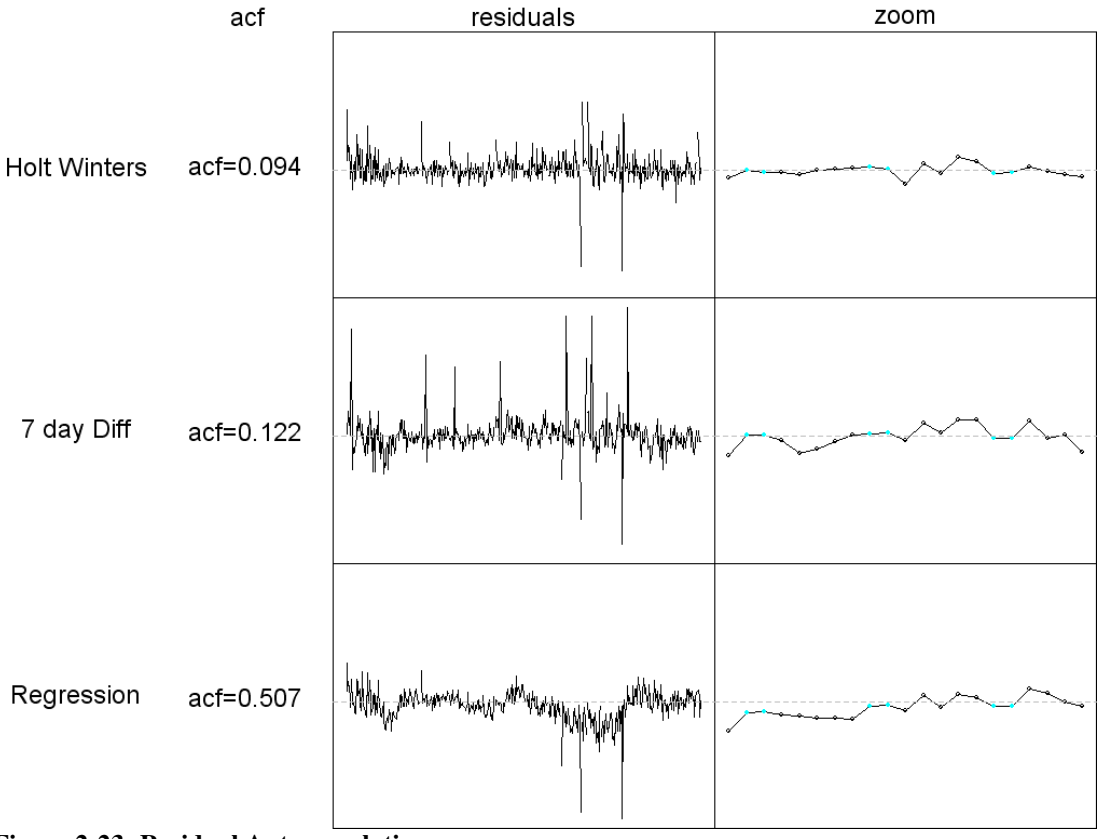

Figure 2-23: Residual Autocorrelation
Residual autocorrelation of the three forecast methods. Y-axes are the same for all graphs. The plots show the overall residuals for each forecasting method and a zoomed-in portion to show daily detail.

In short, the effect of forecast precision on detection performance for these health data is close to expected performance; more precise forecast methods result in improved detection, accounting for seasonal variance improves performance

estimation, and the amount of difference between forecast methods depends on outbreak size.

2.8. Conclusions

In this chapter, we have shown that improved forecasting results in improved detection, both in terms of probabilities of true alert and in timeliness. We examined the effect of forecast precision on detection performance theoretically and quantified the effects under standard control chart assumptions. We have also examined the effects of assumption violation on this relationship, showing that improved forecasting does not always result in improved detection, as in the case of seasonal variance. And in some scenarios although improved forecasting does result in improved detection, the improvement is marginal and might be considered practically marginal (especially when considering the cost of using a more precise forecaster). We conclude that forecasting should be tuned to best capture the background non-outbreak behavior, while detection should be tuned to the outbreak signal. However, the level of investment in more precise forecasts should be weighed against factors such as the required outbreak size, amount of residual autocorrelation, and risks of the forecast method capturing the outbreak.

Several questions arise for practical consideration. First, while we have explored the effects of autocorrelation and seasonal variance, we have not explored the effects of biased or non-normal residuals aside from providing Chebyshev bounds. As we have seen in the authentic data, biases can arise in actual residuals and can affect performance. Second, one additional bias which has not been considered is the effect

of holidays, days with extremely low values that are not predicted by the forecaster. One method for dealing with such cases is described in Section 4.3, but it is also important to have a theoretical understanding for its effect. In addition, while we have examined the detection performance for spike outbreaks and timeliness performance for step outbreaks, a complete delay distribution would include both metrics and give a more complete picture; it would also be relevant to consider average and complete delay distributions for other outbreak shapes, such as exponential or lognormal rise. Some work on this more complete delay picture can be found in Section 5.3.2. Lastly, we have not considered the quality of the training data used for prediction. Not only should it be possible to apply previous work to give expected performance based on the amount of training data (such as the multiplicative Holt-Winters accuracy bound given by (Chatfield & Yar, 1991)), but the impact of outbreaks contaminating the training data or different guardband widths should also be considered.

In conclusion, given the forecasting precision needed for useful detection, the question is whether that level of precision is achievable. This raises the question of whether there is enough quality of signal in pre-diagnostic data. The random elements in the data impose a limit on how well we can forecast, how low an RMSE we can achieve, and ultimately on how well we can detect. It may be that, due to the high noise in most pre-diagnostic data, relatively high false alert rates are required in order to detect outbreaks in a timely manner. For example, if the desired performance is to have a false alert once every two weeks, and have a 95% chance of detecting a

spike outbreak impacting 100 people, to achieve this one would need normal residuals with a forecast RMSE < 32. In contrast, the best forecast method used here has RMSE = 59 on actual data. If we cannot accept a higher false alert rate, then we must either find a way to further improve our forecast methods (e.g., by incorporating other sources of information or by using ensembles), or tailor our detectors to specific outbreak signals.

Chapter 3 : Improved Forecasting Methods

3.1. Introduction

Modern biosurveillance relies on multiple sources of both pre-diagnostic and diagnostic data, updated daily, to discover disease outbreaks. Intrinsic to this effort is the assumption that the data being analyzed contain early indicators of a disease outbreak. However, in addition to outbreak indicators, biosurveillance data streams include factors such as day-of-week effects, seasonal effects, autocorrelation, and global trends. These *explainable* factors obscure outbreak events, and their presence in the data violates standard control chart assumptions. Monitoring tools such as Shewhart charts, Cumulative Sum charts, and Exponentially Weighted Moving Average control charts will alert largely based on these explainable factors instead of on outbreaks. A popular solution is therefore to remove explainable factors from a series, thereby obtaining a series of residuals which do not contain explainable factors. Obtaining such residuals is typically done by forecasting the expected level based on the *explainable* factors. The forecast residuals should then be composed of outbreak signals (if they exist) and a smaller degree of variation, making outbreak signals easier to detect.

By evaluating the residuals from a forecaster in terms of their RMSE, ACF, and Day-of-Week Seasonal Variance, we can estimate their performance using the methods presented in Chapter 2. In this chapter, we first describe some common existing forecasting methods and compare them on these metrics, then present, develop, and evaluate new methods for forecasting.

As discussed in Chapter 2, these residual patterns can negatively affect the performance of anomaly detection from forecast residuals. This can affect the results even to the point of making a less accurate but more well-behaved forecaster be better than a more accurate forecaster with less well-behaved residuals. Because they can have such a dramatic impact on the quality of the resulting control chart performance, determining the most effective method for removing these patterns from a given data set is very important. The tools used in this section show quantitative and qualitative methods for comparing methods' applicability to a syndromic data series and effectiveness at generating residuals with low RMSE, ACF, and Day-of-Week Seasonal Variance.

We use those same tools to evaluate ways of improving forecasting methods, including using cross-series covariates; using additional temperature information; and combining multiple forecasters into an ensemble forecast.

While we consider only forecast methods here, more general preconditioning methods can also be used to remove some of these explainable effects by more advanced methods. Graphical methods to analyze a data set, such as those used in Section 1.3, can also examine the resulting preconditioned data set for its adherence to the detection assumptions (Lotze et al., 2008). Such methods may also be incorporated into the detection; one such method is described in Section 4.3.

3.2. *Current Forecasting Methods*

There are a number of forecasting methods which are in use in modern biosurveillance. These include model-based methods, which assume a particular model and estimate the parameters in that model, and data-driven methods, which fit the data non-parametrically rather than attempting to model the causes. The methods can also differ in their global versus local nature. Here, we discuss the most common methods.

3.2.1. Linear regression models

Regression models are a popular method for capturing recurring patterns such as day-of-week, seasonality, and trends (Rice, 1995). The classic assumption is that these patterns do not change over time, and therefore the entire series can be used to estimate them. To model the different patterns, suitable predictors are created:

Day-of-week effects can be captured by six dummy variables, each representing one day of the week (relative to the remaining baseline day). If there is only a weekday/weekend effect, a single dummy variable can be used.

A global linear trend can be modeled using a predictor t that is a running index (t=1,2,3,…). Other types of trends such as exponential and quadratic trends can also be captured via a linear model by transforming the response and/or index predictor, or by adding transformations of the index predictor (such as adding t^2 to capture a quadratic trend).

Seasonality is most frequently modeled by a sinusoidal trend. The CDC uses a regression model that includes sine and cosine functions to capture a cyclical

trend of mortality rates due to influenza (Serfling, 1963, CDC, 2006), although these terms will not be significant in series without pronounced seasonality. Another regression-based method for dealing with seasonality is to fit local regression models, using past data from the same time of year (Farrington et al., 1996). Note that explicit modeling of seasonal variation assumes that the seasonal pattern remains constant from year to year.

Holidays can be captured by constructing a dummy variable for holidays or by treating holiday days as missing values.

From our experience as well as other reports in the literature (Brillman et al., 2005, Burkom et al., 2007), we find that seasonality effects tend to be multiplicative rather than additive with respect to the response variable. Thus, a linear model where the response is transformed into a natural log ($\ln(y)$) is often appropriate. The regression estimate for a day is transformed back to the original scale to create the forecast. For our data series, we fit a linear regression and a multiplicative regression, and found that the multiplicative version better captured the day-of-week effect. Both are reported below.

Currently, several biosurveillance systems implement some variation of a regression model. ESSENCE uses a linear regression model that includes day-of-week, holiday, and post-holiday indicators (Marsden-Haug et al., 2007) and BioSense uses a Poisson regression with predictors that include a linear trend, sine and cosine effects for seasonality, month indicators, DOW indicators and holiday and day-after holiday indicators (Bradley et al., 2005).

The regression model for our data includes daily dummy variables (*Monday, Tuesday, Thursday, Friday, Saturday, Sunday*) to account for the DOW effect, a holiday indicator (*Holiday*), an index variable (*index*) to capture a linear trend, daily average temperatures (*Tavg*, a method described in Section 3.5) and monthly dummy variables (*Jan, Feb, Mar, Apr, May, Jul, Aug, Sep, Oct, Nov, Dec*) to remove seasonality.

The main advantage of regression modeling is that it provides a general yet powerful method to remove variation due to factors unrelated to outbreaks. It is relatively effective at removing both yearly seasonality and day-of-week variation. However, it requires a fairly large amount of data for obtaining accurate estimates, especially for long-term patterns. Regression is most effective when its assumptions are met: in this case, when the relationship between the predictors (such as day-of-week or year) are consistent over time. While the day-of-week patterns are fairly stable (outside of holidays), some annual patterns can significantly fluctuate over time. In particular, if influenza is not intended to be detected by the system, the timing of influenza's initial growth and its scale of impact are not consistent from year to year, and so its significant impact is difficult to model using regression.

3.2.2. Differencing

Differencing is the operation of subtracting a previous value from a current one. The order of differencing gives the vicinity between the two values: an order 1

differencing means that we take differences between consecutive days ($y_t - y_{t-1}$), whereas an order 7 differencing means subtracting the value of the same day last week ($y_t - y_{t-7}$). This is a popular method in time series analysis, where the goal is to bring a non-stationary time series closer to stationarity (Brockwell & Davis, 1987). Differencing has an effect both on removing linear trends as well as removing recurring cyclic components. In the context of syndromic data, the first instance where differencing was suggested is in (Muscatello, 2004).

In biosurveillance data, the DOW effect can be best accounted for by using an order 7 difference. The forecast is simply the value from 7 days ago, and the residual is simply the difference between the value on the current day and the value 7 days ago. In addition, we explored accounting for holidays by removing the values on holidays, and then obtaining differenced values for the 7th day following a holiday by differencing at lag 14 (i.e., subtracting the value from two weeks prior). This improves the method by removing outliers from known (holiday) causes.

The main advantage of differencing is that it is easy and computationally cheap to perform, and so provides an excellent basis for comparison. It is very effective at removing both weekly and monthly patterns but can result in abnormally high results after abnormally low points in the original data (called "negative singularities" by (Zhang et al., 2003)). Another side-effect of seven-day differencing is that it creates strong weekly partial autocorrelation effects and can increase the variance in the data if there is little or no existing DOW effect.

3.2.3. Holt-Winters exponential smoothing

The Holt-Winters exponential smoothing technique is a form of smoothing in which a time series at time t is assumed to consist of four components: a level term L_t, a trend term T_t, a seasonality term S_t and noise. The k-step ahead forecast is given by

$$\hat{y}_{t+k} = (L_t + kT_t)S_{t+k-M}, \qquad \text{(Eq. 3-1)}$$

where M is the number of seasons in a cycle (e.g., for a weekly periodicity M=7). The three components L_t, T_t, and S_t are updated, as new data arrive, as follows:

$$\begin{aligned} L_t &= \alpha\frac{Y_t}{S_{t-m}} + (1-\alpha)(L_{t-1} + T_{t-1}) \\ T_t &= \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \\ S_t &= \gamma\frac{Y_t}{L_t} + (1-\gamma)(S_{t-M}), \end{aligned} \qquad \text{(Eq. 3-2)}$$

where $\alpha, \beta,$ and γ are smoothing constants that take values in [0,1]. Each component is updated at every time step, based on the actual value at time t. The components are initialized as $L_1 = 0$, $T_1 = 0$, and $S_1 = S_2 = ... = S_M = 1$.

For our data we use the multiplicative seasonality version because the seasonal effects in our syndromic time series are generally proportional to the level L_t. An additive formulation is also available (Chatfield, 1978, Holt, 1957).

The principal advantage of this technique is that it is data-driven and highly automatable. The user need only specify the cycle of the seasonal pattern (e.g., weekly), and the three smoothing parameters. The choice of smoothing parameters depends on the nature of the data and the degree to which the patterns are local versus

global. A study by (Burkom et al., 2007) considered a variety of city-level time series, both with and without seasonal effects. They recommend using the smoothing coefficients $\alpha = 0.4, \beta = 0$, and $\gamma = 0.15$ for seasonal series and $\alpha = 0.1, \beta = 0, \gamma = 0.15$ for series without yearly seasonality. Following this guideline, we used the first settings for each series that exhibited a one-year autocorrelation higher than 0.15 (since a series with yearly seasonality will significantly correlate with itself at one year intervals), and the second setting otherwise. In addition, we applied the modification suggested in (Burkom et al., 2007), which does not update the parameters if the actual value deviates from the prediction by more than 50% (to avoid the influence of outliers).

The Holt-Winters method is very effective at capturing yearly seasonality and weekly patterns. Although it is not straightforward to tune the smoothing parameters, the settings provided here proved generally effective for our syndromic data. One point of caution should be made. As in any method that produces one-step-ahead predictions, a gradually increasing outbreak is likely to get incorporated into the background noise, thereby masking the outbreak signal. One solution is to generate and monitor k-day ahead predictions ($k > 1$) in addition to one-day-ahead predictions.

3.3. Evaluation of Current Forecasting Methods

In this section, we compare the current forecasting methods described in Section 3.2. We use the mathematical foundation from Chapter 2 to perform the analysis. Since

we are using a forecasting method, we can predict its performance using three metrics: the root mean squared error (RMSE), autocorrelation, and seasonal variance. The first two are easily measured. For the third, we determine the residual standard deviation for each day-of-week, and then take the standard deviation of those individual values. While not directly applicable to assessing detection performance, this metric can be used to compare different forecasting methods. Thus, we can create a simple table showing the performance of each method on the various health data. Table 3-1 shows the performance of each method on sales of throat lozenges.

	regression	log_regress	holt-winters	7dayDiff	7dayDiff_holi
RMSE	171.04	176.26	125.53	198.54	171.69
ACF	0.55	0.55	0.14	0.48	0.59
Weekly	23.17	25.23	32.41	37.63	19.09

Table 3-1: Throat Lozenge Forecast Performance Metrics
Forecast performance metrics for throat lozenge sales, comparing five current forecasting methods applied to biosurveillance data.

To show information on multiple series at once, we can also create small-multiples histograms. In these graphs, each histogram shows the distribution of one statistic for one method, over the different series in a data set, as well as printing the mean. For example, Figure 3-2 shows this for the OTC medication sales described in Section 1.3.2. The first column contains the histograms of the method's RMSE over each of the 8 OTC series. Each row has the results for one method. Over the 8 series, the residuals from using a regression forecaster had a mean of 77.64, while regression on the log values had a mean RMSE of 84.30.

Figure 3-1, Figure 3-2, Figure 3-3, and Figure 3-4 show the methods' performance over the different data sets (all 3 authentic data sets combined, OTC medication sales,

ED visits, and BioALIRT, respectively). From these comparisons, it is clear that Holt-Winters consistently outperforms regression and differencing not only in having low RMSE, but also in terms of low autocorrelation and low day-of-week seasonal variance.

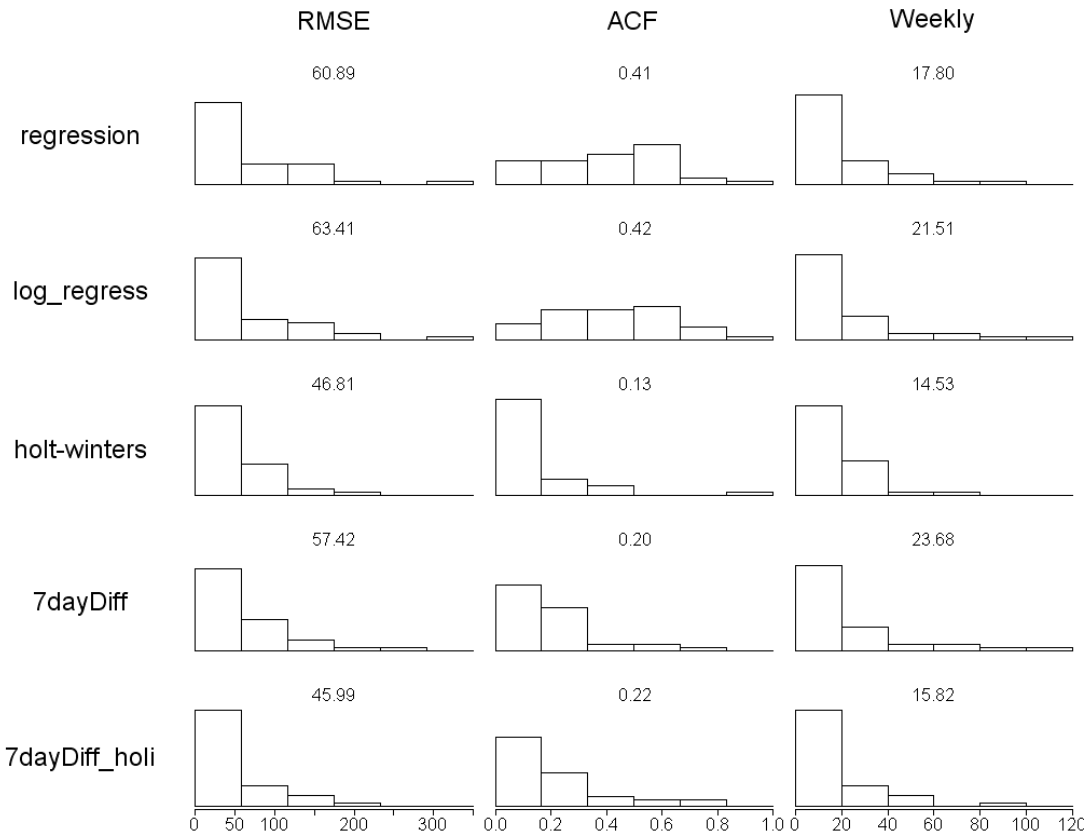

Figure 3-1: Forecasting Comparison Overall
Comparison of different forecasting methods on the OTC, ED, and BioALIRT data sets combined. Each histogram shows the distribution of the residuals for one method, for one metric, across all the data series in the data set. The mean value (over all data series) is also printed above each histogram.

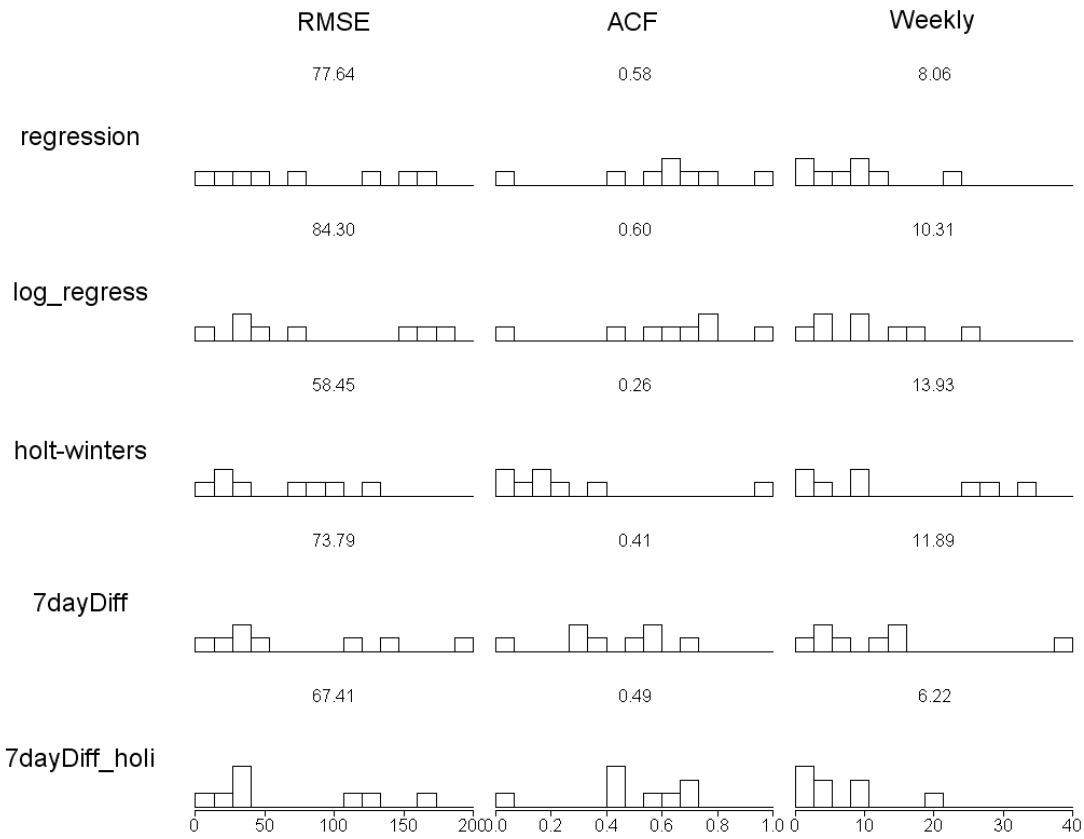

Figure 3-2: Forecasting Comparison for OTC

Comparison of different forecasting methods on the OTC medication sales data set. Each histogram shows the distribution of the residuals for one method, for one metric, across all the data series in the data set. The mean value (over all data series) is also printed above each histogram.

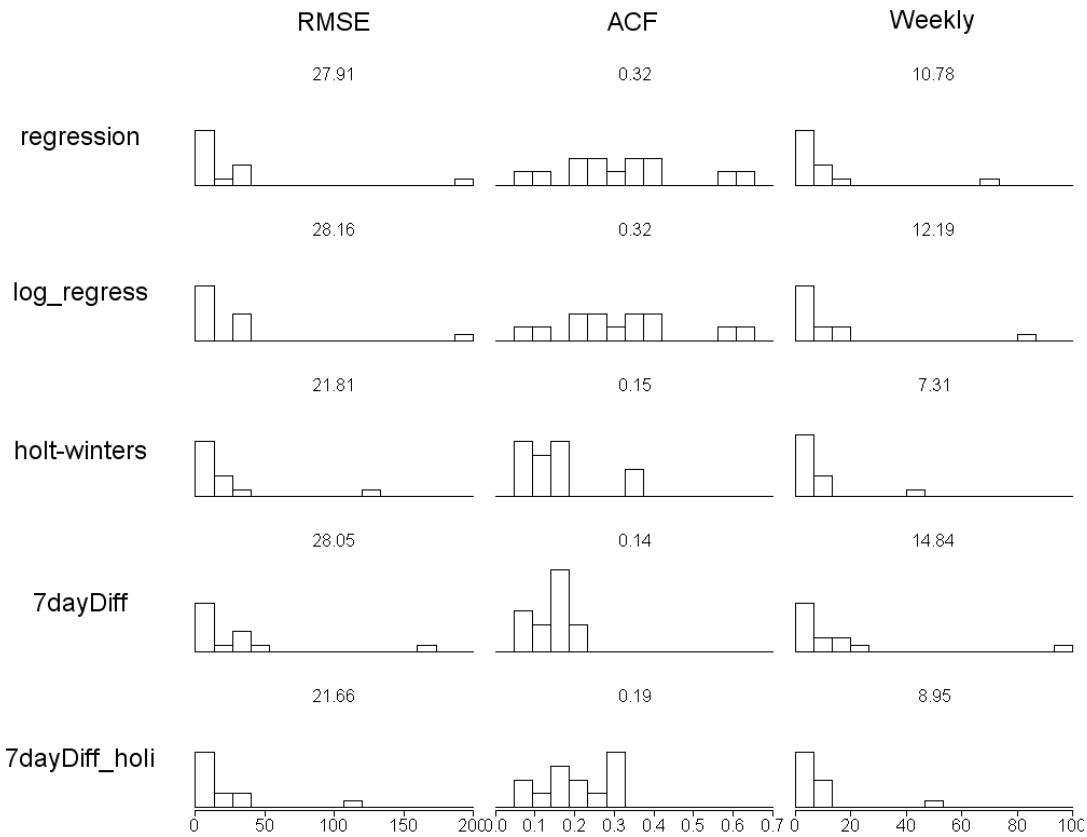

Figure 3-3: Forecasting Comparison for ED

Comparison of different forecasting methods on the ED syndromes data set. Each histogram shows the distribution of the residuals for one method, for one metric, across all the data series in the data set. The mean value (over all data series) is also printed above each histogram.

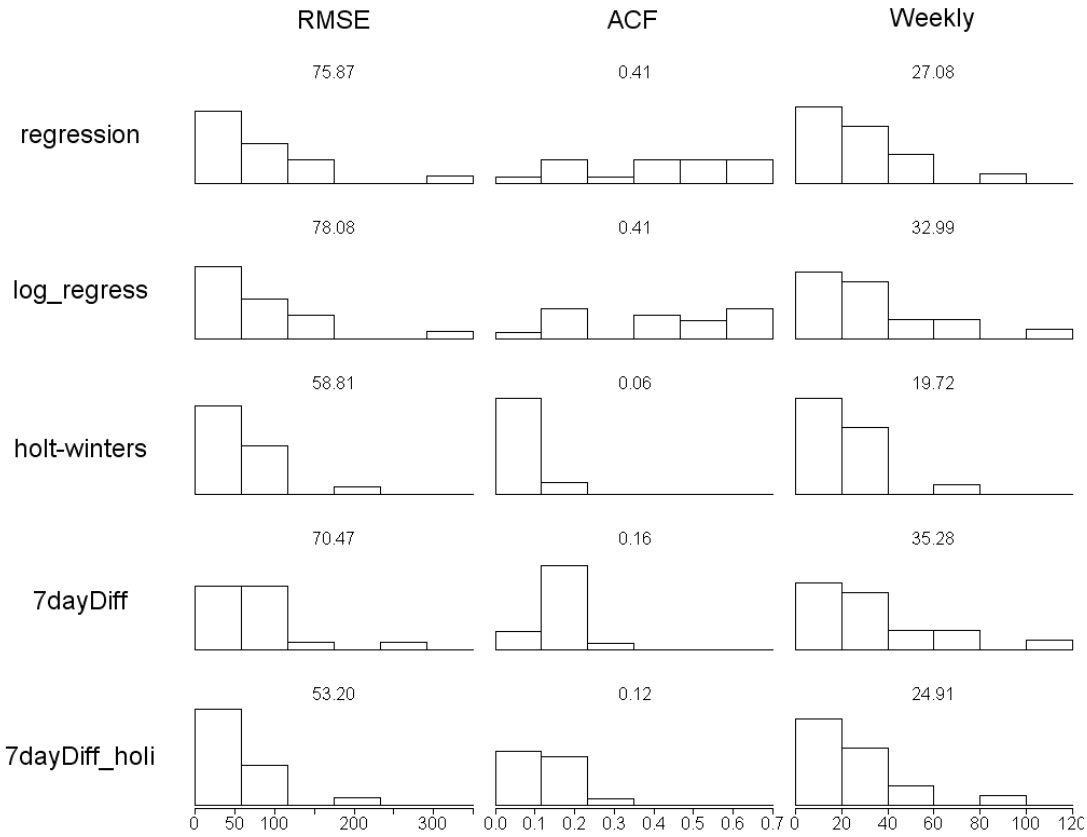

Figure 3-4: Forecasting Comparison for BioALIRT
Comparison of different forecasting methods on the BioALIRT data set. Each histogram shows the distribution of the residuals for one method, for one metric, across all the data series in the data set. The mean value (over all data series) is also printed above each histogram.

3.4. Cross-Series Covariates

In most biosurveillance data, there are a number of indicators which are tracked at the same time. By using this additional data, we can improve the forecasts of the health series we are interested in (presumably the one we expect to be impacted by an outbreak). In over-the-counter purchase data, there are multiple data categories such as throat lozenge sales, headache medicine sales, and liquid decongestant sales. In emergency room counts, there are counts of multiple symptoms. These data will be impacted by many of the same explainable effects which impact the series of interest. In particular, they will both be impacted by effects which are not related to the

outbreak, but which may not be recorded in the other predictor variables. For example, if there is a storewide three-day sale, then during those three days, there will be an increase in OTC sales across all types. This also applies to more consistent and subtle factors, such as the fact that one on-duty nurse may be more efficient at admitting patients than others, resulting in an increased number of patients overall during their shift. Because of this joint influence by explained factors, using other associated health series as predictors can be used to remove underlying factors which are not measured, but which will have an impact on the series of interest.

The main method used here for incorporating information from other relevant series is standard linear regression. Although many other methods are possible, linear regression is a standard first option and should demonstrate whether or not cross-series information can be used to improve prediction. For this reason, a multiple regression prediction method, to be used as the baseline (excluding cross-series information), is performed with the following predictors:

- six dummy variables, each representing one day of the week
- a predictor *t* that is a running index ($t = 1, 2, 3, ...$)
- yearly sine term
- yearly cosine term

The prediction model is applied such that for each day, coefficients are estimated using all previous days, and then a prediction is made for the next day using those estimated coefficients. To determine the improvement due to using information from

other series (hereafter referred to as cross-series covariates), a separate regression model is fit using the same predictors plus additional predictors for each related series from the previous day. For example, when forecasting civilian gastrointestinal visits, two additional predictors are added:

- military gastrointestinal visits from the previous day
- gastrointestinal prescriptions from the previous day

A comparison of cross-series and univariate regression models, and Holt-Winters exponential smoothing is given in Figure 3-5.

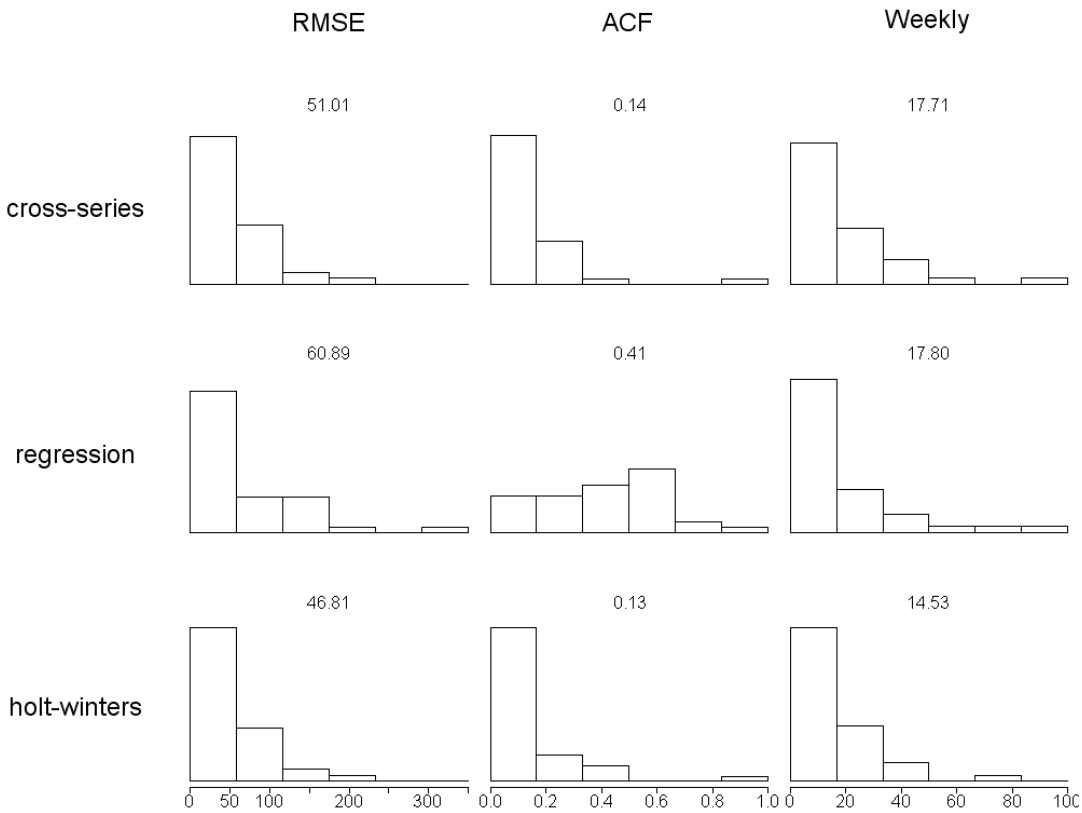

Figure 3-5: Forecast Comparison for Cross-Series Regression
Comparison of forecasts from using cross-series predictors (rather than univariate predictors), on the combined authentic data sets.

The use of cross-series covariates (other series used as predictors) to improve the forecast gives a significant improvement in terms of RMSE, and a striking

improvement in autocorrelation. When using such a technique, one must take care that the outbreak of interest will not occur over all monitored series--in such a case, this could result in diminished performance by effectively filtering out the outbreak. In general, however, by improving the regression techniques, or adapting other methods to utilize covariate series, biosurveillance methods should see an improvement in forecasts and a corresponding improvement in detection.

3.5. Using Temperature as a Predictor

Regression models can also be used to integrate external information that can assist in removing explainable patterns. For example, seasonal patterns tend to be highly correlated with temperature. Figure 3-6, which shows counts of daily respiratory complaints and the average daily temperature, demonstrates this relationship. There is a strong negative relationship between temperature and sales: as the weather gets colder, more cough remedy drugs are sold.

In many cases, daily temperature has a significant relationship with disease. While the link between weather and disease has long been known (the relationship is clear in Figure 3-6), it has only recently been modeled biologically; for example, it has been recently shown (Lowen et al., 2007) that the lower temperatures in winter contribute to increased flu transmission by increasing the amount of time the flu virus can

survive.

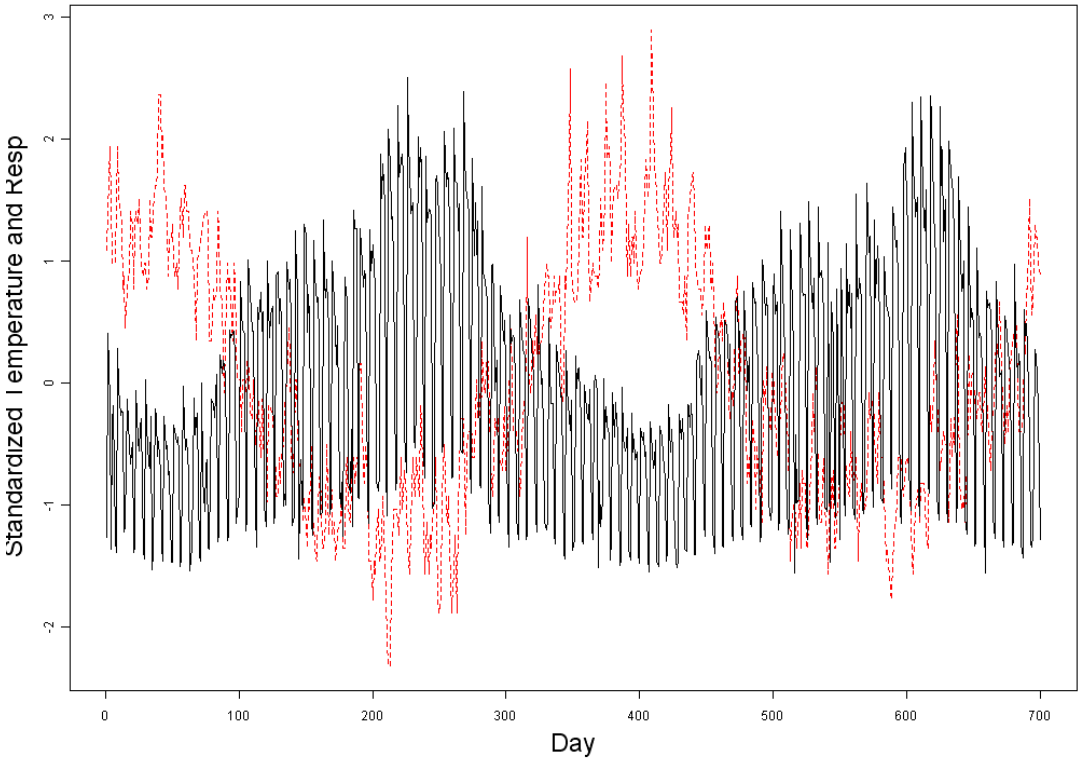

Figure 3-6: Temperature and Respiratory Visits
Timeplot showing counts of daily counts of military respiratory complaints (in black) and the average daily temperature (in dotted red). Both series have been standardized, in order to be plotted on the same graph.

Temperature and other weather data can be extracted from NOAA records, available online from the NOAA National Climate Data Center, http://www.ncdc.noaa.gov/oa/ncdc.html. The comparisons below use NOAA station records as the source of weather data. The use of temperature as a predictor is not found elsewhere in biosurveillance literature, but promises to be useful and relevant in forecasting health series levels. We show a comparison, using a regression forecaster without temperature (using daily dummies, trend index, and yearly sine and

cosine terms) as compared to one which has the average temperature added as an additional predictor. Figure 3-7 shows the results.

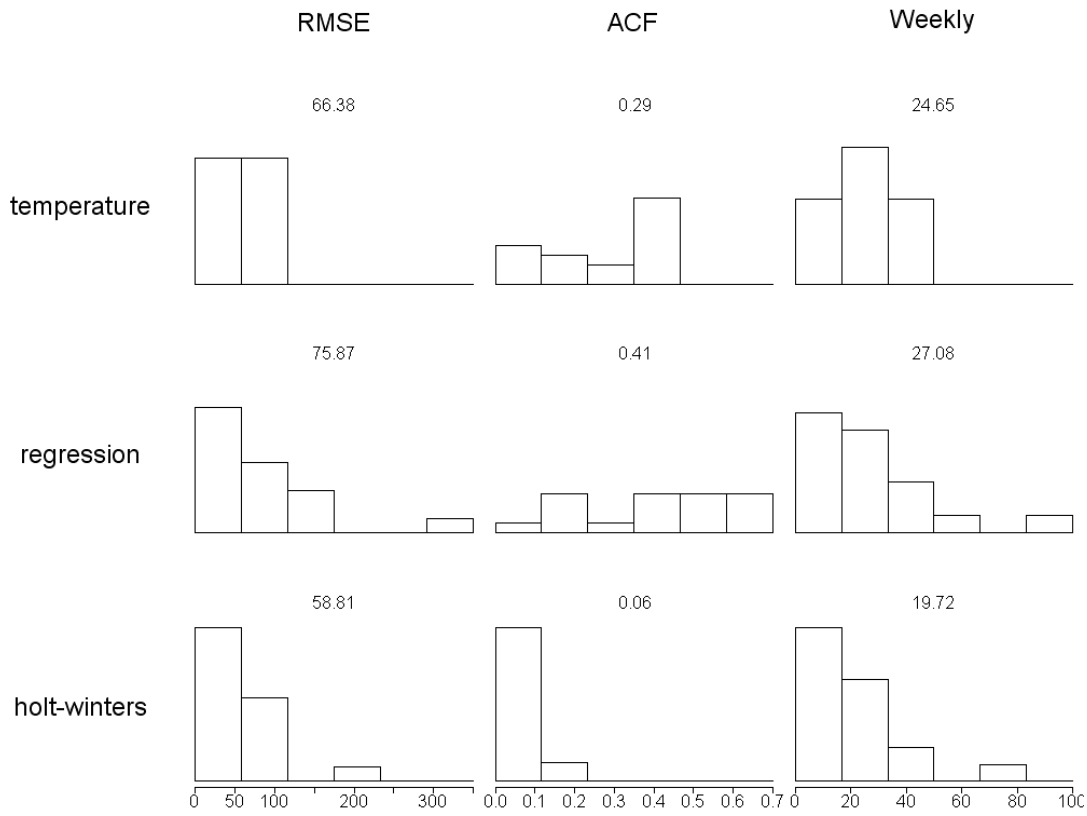

Figure 3-7: Forecast Comparison for Temperature Regression
Comparison of regression with temperature (top), regression without temperature (middle), and Holt-Winters (bottom) forecasters, as applied to the combined authentic data sets.

Temperature clearly has a significant impact on reducing RMSE, autocorrelation, and day-of-week seasonal variance. This is impressive, given that it occurs even when the original regression has sine and cosine terms for seasonality. This shows that temperature is a more tightly correlated predictor with disease than simply an annual pattern. It is worth noting that the addition of temperature alone is not sufficient to make vanilla regression competitive with Holt-Winters for forecasting biosurveillance

data. However, the strength of its improvement indicates that temperature can be used to significantly improve other methods' results as well.

3.6. Ensemble Forecasting for Biosurveillance Data

3.6.1. Ensemble Method

There are many different forecasters available for use with biosurveillance data. None is perfect for all types of biosurveillance data or on all days throughout the year. Because of this, we should be able to take advantage of each individual forecaster's strengths and create a *combined* forecaster which is better than any of the individual forecasting methods. It is this concept which we explore as we develop an ensemble forecaster for biosurveillance data. The work in this section is based on previously published work in (Lotze & Shmueli, 2008a).

Multiple forecasters are generated for each time series. For each day, the linear combination of forecasters which has the minimum squared error on past days is determined; this can be found by running a simple linear regression using the past time series values, with the past forecasts as predictors. The resulting linear combination is used to combine the forecasts, creating an ensemble forecast value for the next day. Residuals are then generated by subtracting the forecast from the observed value for each day.

Specifically, if we have F forecasters, $f^1 \ldots f^F$, each making forecasts on days $1 \ldots t$, where forecaster f^k makes forecasts $f_1^k \ldots f_t^k$, then the ensemble forecaster, for day t,

provides the forecast $f_t^e = \beta_{(0,t)} + \beta_{(1,t)} f_t^1 + \beta_{(2,t)} f_t^2 + ... + \beta_{(F,t)} f_t^F$, where the $\beta_{(i,t)}$ values are chosen to minimize the squared error on past days, $\sum_{i=1}^{t-1} (f_i^e - y_i)^2$. The residual value for day t is then $r_t = f_t^e - y_t$.

As the nature of the series changes over time, each of the forecasters has a different forecast accuracy level. By changing the linear coefficients to reflect this, the ensemble forecaster adapts to take advantage of the local accuracy of different individual forecasters.

3.6.2. Results

For these results, we used three methods: a 7-day difference, a Holt-Winters Exponential Smoother, and a linear regression. The linear regression used as predictors day-of-week dummy variables, cosine and sine seasonality terms, and a linear index term. The analysis was run on three data streams from the ISDS contest (described in Section 1.3.4):

1. Patient emergency room visits (ED) with gastrointestinal symptoms
2. Aggregated over-the-counter (OTC) anti-diarrheal and anti-nauseant sales
3. Nurse advice hotline calls (TH) with respiratory symptoms

Recall that each of these series had five years of non-outbreak data. Forecasting methods were trained on two years of data, and their RMSE tested on the last three. The ensemble method had the lowest RMSE on each series. Results are in Table 3-2.

	ED	OTC	TH
Regression	20.18	113.47	6.08
7-day Diff	23.62	135.01	8.12
Holt-Winters	18.20	110.12	6.38
Ensemble	18.05	103.66	5.94

Table 3-2: Ensemble RMSE Comparison
RMSE for regression, 7-day Diff, Holt-Winters, and ensemble forecasters, on ISDS contest data. Ensemble has the lowest RMSE in all cases.

Figure 3-8 shows the distribution comparison for the ensemble method over the BioALIRT data set. Its results are approximately the same as the best method, Holt-Winters. When there are more methods with comparable performance, the ensemble method should be able to gain further improvement by combining them.

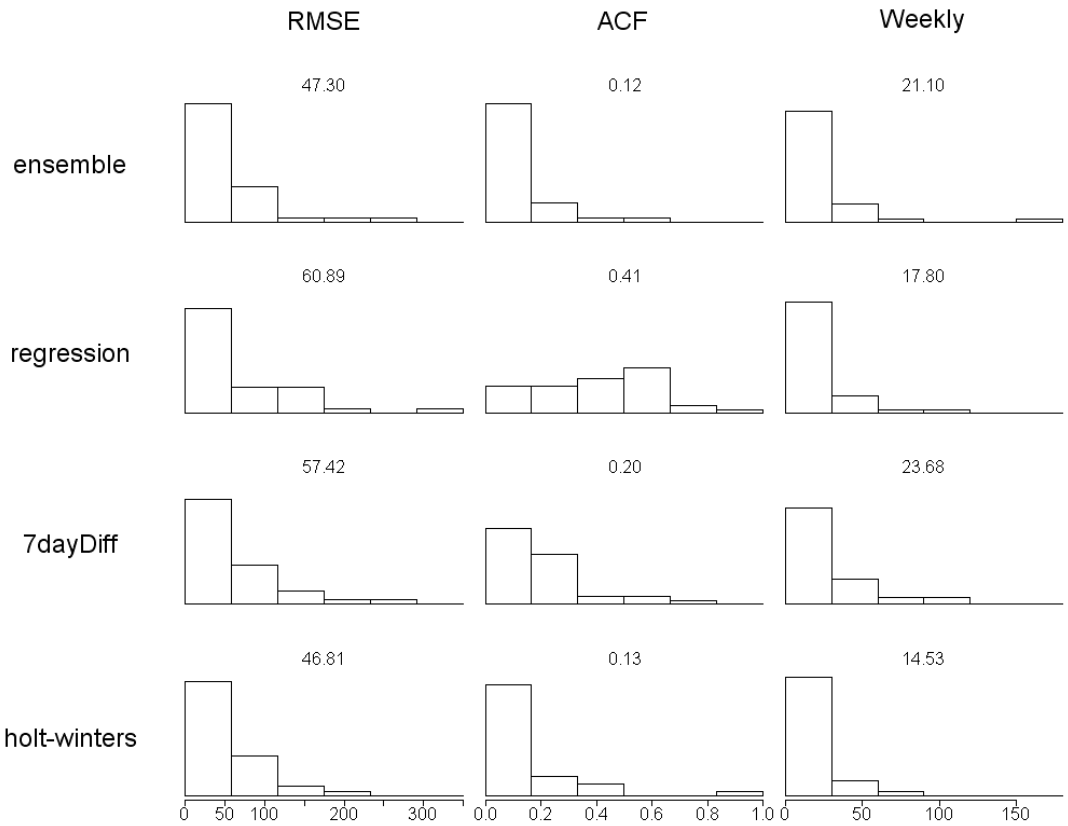

Figure 3-8: Forecast Comparison for Ensemble Forecast
Comparing RMSE, autocorrelation, and weekly variance for ensemble, regression, 7-day diff, and Holt-Winters on the combined authentic data sets.

3.7. Conclusions and Future Work

We have proposed and examined several new methods and improvements to existing methods for forecasting biosurveillance data. All of these methods can be used to improve the forecasting of biosurveillance data, and thus to improve detection. In addition, we have presented a way to display the forecasting results and compare them on forecast accuracy, autocorrelation and day-of-week seasonal variance.

Although we present visualization tools to improve the performance of different forecasters, we also caution that domain expertise will also be necessary to create improved forecasters. For example, the day-of-week effect can often be explained by the fact that many hospitals dramatically reduce staffing on weekends (Tarnow-Mordi et al., 2000, Czaplinski & Diers, 1998, Kovner & Gergen, 1998, Blegen MA, 1998, Strzalka & Havens, 1996, McCloskey, 1998, Archibald et al., 1997), and so counts are generally much lower on weekends. Marketing knowledge can tell us that grocery shopping is more popular on weekends than on weekdays. And for both types, holidays always have exceedingly low counts (except for some areas such as those with high-risk sports); domain expertise can often distinguish between an official holiday and an observed holiday impact. Domain expertise is an invaluable tool for explaining and accounting for explainable patterns in biosurveillance data. Although it is tempting to completely automate the analysis and preprocessing of syndromic data series, human intervention is still a critical part of the solution.

Although we focus here on data that are used in temporal monitoring using control charts, such preprocessing can also be helpful in spatial and spatio-temporal monitoring, when an underlying *iid* assumption exists, such as in the widely-used spatio-temporal scan statistic (Kulldorff, 2001).

There are still a number of forecasting improvements which we have not considered here, but which could be quite promising. For example, robust forecasting methods (which reduce the influence of outliers in the training data) may also be helpful, particularly in dealing with unmarked holidays, past outbreaks in the training data, or other unexpected changes in behavior, such as special sales or otherwise busy shopping days. In addition, while we have evaluated other models such as the COM-Poisson (Shmueli et al., 2005) on health data, but found their performance lacking, further modification of these methods could provide an effective forecaster. In addition, our earlier work on wavelets (Lotze et al., 2006) indicates that their performance is competitive, providing a good potential starting point for improvement. Similarly, we have only briefly touched on the topic of sliding window methods, which can be used to improve forecasting when the data structure changes over time, such as in the case of seasonal covariance.

We are also concerned with the ability of cross-series forecasters to respond to seasonal covariance, since these methods assume a constant relationship between the series. Analysis to determine the robustness of different forecasters to seasonal covariance should be undertaken.

One future direction is to create an automated application that uses these forecasting methods to explore and categorize each data series, providing recommendations and rationales for various methods to the end user. This automated expert system could help practitioners determine the methods which would best forecast their data, while allowing them to include domain knowledge. Such a system could perform this function by analyzing the statistics above, selecting appropriate forecasting methods, and then displaying graphical plots to illustrate the reasons for the each suggested method. The user would then be able to assess which patterns are reasonable in a particular data set, and based on the system's output, to choose the preferred forecasting operation(s).

In our analysis, we assumed that there were no known outbreaks in the baseline data. However, it is obvious that the data contain seasons of influenza which affect both ED visits and OTC sales. The problem of unlabeled data, in the sense that we do not know exactly when a disease outbreak is present and when there is no disease, is a serious one for both modeling and performance evaluation. A related issue that arises in monitoring daily data is that of gradual outbreaks. Autocorrelation between days (in particular, 1-day autocorrelation) should also be examined and controlled for, in order to approach the statistical independence assumption required for standard control charts. However, a gradual outbreak will also increase the autocorrelation between days (as a rising number of people will show symptoms). It is therefore important to remember the danger of embedding the outbreak signal into the

background data. As proposed earlier, one solution is to examine predictions that are farther into the future, and also to use a "guard band" that avoids the use of the last few days in the detection algorithm (Burkom et al., 2004).

Chapter 4 : Improved Detection Methods

4.1. Introduction

In Chapter 1, we described the purpose of biosurveillance as detecting disease outbreaks in a timely manner with few false alerts. We also defined the metrics for evaluating detection. In Chapter 2, we showed how improved forecasting could result in improved detection, and in Chapter 3, we described some methods to improve forecasting of baseline biosurveillance health data. In this chapter we describe methods that are aimed at improving the detection step that follows the forecasting step. We present three categories of methods for improved detection. The first category (in Section 4.2) consists of three related methods for improving detection when one has access to multiple series of health data; the outbreak signal may appear in either individual or multiple series. The second category (in Section 4.3) considers post-processing techniques to deal with the day-of-week seasonal variance issues presented in Chapter 2. Finally, Section 4.4 examines a fourth category of methods, which are based on optimizing detection of specific outbreak patterns; this is useful when the type of outbreak of interest and its signature in the data are specified.

4.2. Multivariate Outbreak Methods

4.2.1. Combination Methods

When a disease outbreak appears in multiple series, combination methods can provide an improved way to detect such outbreaks. Combination methods are used to combine multivariate series measuring the same syndrome into a single univariate series to measure that syndrome. This is done in order to reduce the variance of the

series and improve the strength of the syndromic signal component relative to the variance due to noise.

As a simple example of this idea, consider the case where there are J series $Y_1 \ldots Y_J$. Each is an independent noisy measure of the underlying signal X_i with identical independent variance, according to the model $Y_{ij} = X_i + \epsilon_{ij}$, where $E(\epsilon_{ij}) = 0, V(\epsilon_{ij}) = \sigma^2$. Then by taking the mean of these series, we can reduce the variance. If $U_i = \bar{Y_i} = \sum_{j=1}^{J} Y_{ij}/J$, then this is an unbiased estimator of the underlying syndromic signal X_i with lower variance than any of the individual univariate series (each univariate series has variance σ^2, but the combined series has variance σ^2/J and standard deviation $\sigma/\sqrt{J}$).

One can see that different methods of combining these series will have different effects depending on the expected outbreak signal. While the variance of the combined series will be reduced, if the outbreak appears in only one of the J data series, then the resulting increase in $\bar{Y_i}$ will be only $1/J$ times its size in the single series. Thus, we will have only a slight increase in our chance to detect it. (Although the outbreak signal is reduced to $1/J$ its size, this must be compared to the standard deviation $\sigma/\sqrt{J}$. The formulas in Chapter 2 can be used to calculate the improvement, which depends on the false alert level as well as the size of the outbreak relative to the square root of the number of series.)

If the outbreak occurs in all series, then its appearance relative to the standard deviation will be J times its size in any individual series, so we will have a much improved chance of detection. In order to determine the effectiveness of a method, we compare its performance to other methods on the same data; we will do this when the outbreak occurs in single series as well as in all series. A natural comparison for any multivariate combination method is against multiple univariate tests, where each series is tested separately.

There has been some scattered work done on multivariate methods in biosurveillance, but the area is largely incomplete. (Burkom et al., 2004) analyzes several multivariate and multiple univariate methods on the DARPA BioALIRT data described in Section 1.3.1. However, it stops short of comparing the performance over multiple types of outbreaks, to determine when one method would be preferred over another. A working paper (Yahav & Shmueli, 2007) compares Hotelling's T^2 with directionally sensitive multivariate EWMA (MEWMA) and multivariate CuSum (MCUSUM) methods in terms of their robustness to assumption violations. While it shows the change in alert rate due to increasing the number of monitored series, it does not provide direct comparisons when the actual false alert rate is held the same between methods. Finally, (Fricker et al., 2008a) compared the directionally sensitive MCUSUM and MEWMA over several forms of baseline data and outbreaks, but did not compare them to multiple univariate methods or identify the performance in terms of the number of series where the outbreak occurs. The work in this section is similar to these past efforts, but directly compares combined multivariate detection with

multiple univariate detection over a variety of factors, controlling for false alert rate. It also introduces some new methods for analyzing multivariate series and compares their effectiveness. Finally, it compares the effectiveness of preconditioning at two different possible points in the multivariate combination.

We consider three different methods of combining multivariate measures into a single measure: standardized mean, principal components analysis, and Mahalanobis distance. We describe each of these next.

4.2.1.1. Standardized Mean

To create the standardized mean, one takes each individual series and standardizes it by subtracting the sample mean and dividing by the sample standard deviation (for that series). In other words,

$$N_{ij} = (Y_{ij} - \hat{\mu}_j)/\hat{\sigma}_j. \qquad \text{(Eq. 4-1)}$$

To do this in an adaptive way, one subtracts the moving average from the past z days and divides by the sample standard deviation from the last z days. We use $z = 56$ as a reasonable length of recent data (2 months) which is also a multiple of 7, and so includes the same number of days-of-the-week.

$$N_{ij} = (Y_{ij} - \hat{\mu}_{j,i-1:i-56})/\hat{\sigma}_{j,i-1:i-56} \qquad \text{(Eq. 4-2)}$$

Then the combined series equals the simple mean of the J standardized series,

$$U_i = \bar{N}_i = \sum_{j=1}^{J} N_{ij}/J. \qquad \text{(Eq. 4-3)}$$

In the tabled results, we refer to this method as 'Normsum', the sum of normalized variables.

4.2.1.2. PCA

Principal Components Analysis (PCA) is a common method for reducing the dimension of multivariate data to a smaller number of variables (Jobson, 1992). It converts the multivariate data into a new basis, in which the first component has the greatest variance of any linear combination, the second has the greatest variance of any linear combination which is orthogonal to the first; and all remaining components have the greatest variance of any combination which is orthogonal to all previous combinations. By ignoring later combinations (those with low variance), one can find orthogonal linear combinations of the variables which capture most of the variance of the original data, but often in far fewer variables.

In general, and in our analysis, the correlation method is used. The sample correlation matrix, R, is found from the sample data. Then the principal components are simply the eigenvectors of R. As a combination method, one can take the first principal component as U, i.e.,

$$U_i = e_1' X_i \qquad \text{(Eq. 4-4)}$$

where e_1 is the eigenvector of R with the largest eigenvalue.

To do this in an adaptive way, one simply uses the correlation matrix from the previous 56 days, using $R_{i-1:i-56}$ instead of R.

4.2.1.3. Mahalanobis Distance (T^2)

The Mahalanobis distance is the standard method for computing distance in a multivariate space, while accounting for the covariance structure of multivariate data;

it is also the basis for the multivariate T^2 test. In the following we first describe how the Mahalanobis distance is used to standardize multivariate observations, and then we show how we use this standardization for combining a multivariate series into a single series.

To normalize multivariate data using the Mahalanobis distance, the mean vector is subtracted from the data vector, and then the result is multiplied by the square root of the inverse of the covariance matrix, $\Sigma^{-1/2}$. Therefore, if X_t is a multivariate set of observations on day t, mean μ and covariance Σ, then the standardized data, given by $Z_t = \Sigma^{-1/2}(X_t - \mu)$, have covariance I. This standardization is therefore a way to create uncorrelated random variables with mean 0 and standard deviation 1 from multivariate data (Rencher, 2002). If the data are multivariate normal, then the resulting variables z_{it} in the normalized vector Z_t are uncorrelated standard normal variables.

Therefore, their squared sum, $\sum_i z_{it}^2 = (X_t - \mu)'\Sigma^{-1}(X_t - \mu)$ follows a χ^2 distribution. In the multivariate normal case where the means and standard deviations are unknown, but estimated from the data, we must instead compute an estimated normalized vector, $T_t = S^{-1/2}(X_t - \bar{X})$. Because in this sum, the mean and variances are estimated, the sum $D_t^2 = (X_t - \bar{X})'S^{-1}(X_t - \bar{X})$ follows a T^2 distribution.

To use the Mahalanobis distance as a combination method, we estimate the covariance matrix Σ and mean μ from the multivariate data using standard estimates S and $\bar{X}$. Then for a single day's multivariate observation vector X_t, one computes the vector $T_t = S^{-1/2}(X_t - \bar{X})$ and then sums the elements of the vector to obtain a univariate series

$$\mathrm{U}_t = \sum t_{it} = (1, 1, ..., 1)S^{-1/2}(X_t - \bar{X}). \qquad \text{(Eq. 4-5)}$$

One could also use the sum of squares, $D_t^2 = (X_t - \bar{X})' S^{-1}(X_t - \bar{X})$ to get the T^2 statistic, and use it to test whether the multivariate vector for day t lies in an appropriate region of the vector space (where the original series are greater than expected). We do not examine this testing approach here, but it is described and analyzed in (Yahav & Shmueli, 2007).

To compute the Mahalanobis-based combination series in an adaptive way, one simply uses the covariance matrix from the previous 56 days, using $S_{i-1:i-56}$ instead of S: $A_t = \sum a_{it} = (1, 1, ..., 1)S_{i-1:i-56}^{-1/2}(X_t - \bar{X}_{i-1:i-56})$.

4.2.2. Empirical Performance Comparison

In order to determine which combination method to use, we must compare their ability to detect outbreaks. However, the detection performance may depend significantly on whether the outbreak shows up in only one of the monitored series, or in all. In other words, performance may depend on how well the series have been chosen to reflect the disease impact. We perform this test by analyzing the results of

simulated outbreaks inserted into the authentic BioALIRT data described in Section 1.3.1.

The detection performance is estimated by comparing performance on two types of single-day ("spike") outbreaks. In one, the outbreak signal appears in each series for the given syndrome simultaneously (where a syndrome is either gastrointestinal or respiratory). In the other, it appears only in one series. Thus, there are 8 possible signal types (6 for the individual series and 2 for the two types of simultaneous syndrome outbreak). Although in practice the outbreak could occur in a subset of some series rather than one or all (even with different sizes), this comparison will serve to illustrate the difference between univariate and cross-series detectors.

To compare detection performance for a specific outbreak type, simulated outbreaks are created for each possible day after the first 56 days (to provide ramp-up time). The size of the outbreak is set equal to one standard deviation of the baseline health series. Each method is run on each outbreak injection, and a ROC curve is generated. Four metrics are then generated: Detection Rate at a fixed false alarm rate of once every 28 days; Detection Rate at a fixed false alarm rate of once every 56 days; Detection Rate at a fixed false alarm rate of once every 72 days; and integrated Detection Rate (area under the ROC curve) for false alarm rates between every 14 to 112 days.

For each combination method, we also consider the addition of a preconditioning step to remove explainable patterns. One can consider preconditioning each univariate series separately first, then combining the preconditioned series; or one could combine the raw data first, then apply a preconditioning method to the combined series. While both options were tested, it turns out that preconditioning the already-combined series is not effective; therefore, we only report results from preconditioning, then combining the preconditioned series in our performance comparison.

In the following, results are reported for detection performance of the various methods. The performance statistics are the Detection Rate at a three fixed false alarm rates (once per 28 days, once per 56 days, and once per 72 days) as well as the proportion (from 0 to 1) of the possible detection area between false alarm rates of once every 14 days to once every 112 days. We compared the different combination methods to three alternatives in terms of detection:

- Simple univariate Shewhart detection, applied to single series' residuals from Holt-Winters (HW) forecasting (described in 3.1.2.4). Results are displayed in columns labeled uniHW or uniHW_(series name)
- Simple univariate Shewhart detection, applied to single series' residuals from univariate regression forecasting (described in 3.1.2.1). Results are displayed in columns labeled uniReg_(series name).
- Simple univariate Shewhart detection, applied to single series' residuals from multivariate regression forecasting on a single target series (described in 3.3.2).

Results are displayed in columns labeled multiReg_(series name)).

We also examined a version of all regression models and combination methods where a univariate HW was initially applied to each individual series before applying the method, for purposes of preconditioning. Results for this version are denoted with the suffix "HW".

4.2.2.1. Single Series Outbreak

In this section, we report the detection results from inserting an outbreak signal into one series at a time.

respPrescrip

	uniHW	uniReg	multiReg	mahalanobisHW	normsumHW	PCAHW
at28	0.026	0.034	0.030	0.040	0.039	0.037
at56	0.009	0.017	0.017	0.022	0.022	0.019
at72	0.009	0.014	0.014	0.014	0.020	0.014
in14_112	0.031	0.035	0.039	0.044	0.046	0.040

respMilVisit

	uniHW	uniReg	multiReg	mahalanobisHW	normsumHW	PCAHW
at28	0.031	0.034	0.043	0.042	0.039	0.037
at56	0.012	0.019	0.017	0.022	0.022	0.019
at72	0.011	0.016	0.016	0.016	0.020	0.014
in14_112	0.036	0.039	0.043	0.045	0.046	0.040

respCivVisit

	uniHW	uniReg	multiReg	mahalanobisHW	normsumHW	PCAHW
at28	0.031	0.034	0.039	0.042	0.039	0.036
at56	0.011	0.017	0.016	0.022	0.022	0.019
at72	0.006	0.011	0.011	0.016	0.020	0.014
in14_112	0.034	0.040	0.044	0.045	0.046	0.040

Table 4-1: Individual Series Outbreak Detection Rates (Resp)
Detection rate for an outbreak inserted into each individual series. Recall that the metrics for evaluation are Detection Rate at a fixed false alarm rate of once every 28 days; Detection Rate at a fixed false alarm rate of once every 56 days; Detection Rate at a fixed false alarm rate of once every 72 days; and integrated Detection Rate (area under the ROC curve) for false alarm rates between every 14 to 112 days. Normsum

refers to the Standardized Mean described in 4.2.1.1. The best performance in each row is shaded.

giPrescrip

	uniHW	uniReg	multiReg	mahalanobisHW	normsumHW	PCAHW
at28	0.028	0.039	0.039	0.040	0.043	0.028
at56	0.009	0.019	0.022	0.019	0.016	0.012
at72	0.006	0.012	0.017	0.016	0.012	0.008
in14_112	0.036	0.043	0.041	0.043	0.040	0.036

giMilVisit

	uniHW	uniReg	multiReg	mahalanobisHW	normsumHW	PCAHW
at28	0.026	0.037	0.040	0.047	0.043	0.028
at56	0.020	0.020	0.025	0.023	0.016	0.012
at72	0.011	0.014	0.016	0.017	0.012	0.008
in14_112	0.038	0.047	0.046	0.048	0.042	0.036

giCivVisit

	uniHW	uniReg	multiReg	mahalanobisHW	normsumHW	PCAHW
at28	0.040	0.047	0.045	0.043	0.043	0.028
at56	0.011	0.020	0.023	0.020	0.016	0.012
at72	0.009	0.019	0.017	0.017	0.012	0.006
in14_112	0.039	0.046	0.049	0.045	0.041	0.035

Table 4-2: Individual Series Outbreak Detection Rates (GI)

Detection rate for an outbreak inserted into each individual series. Recall that the metrics for evaluation are Detection Rate at a fixed false alarm rate of once every 28 days; Detection Rate at a fixed false alarm rate of once every 56 days; Detection Rate at a fixed false alarm rate of once every 72 days; and integrated Detection Rate (area under the ROC curve) for false alarm rates between every 14 to 112 days. Normsum refers to the Standardized Mean described in 4.2.1.1. The best performance in each row is shaded.

From the ROC curve statistics in Table 4-1 and Table 4-2, it seems clear that Mahalanobis and Normsum are providing improved performance. It is particularly striking that even when the outbreak is inserted into only one series, the multivariate methods use the extra information well enough to provide comparable performance to the univariate methods, and often improved performance.

4.2.2.2. All-series Outbreak

In this section, the detection results are for an insertion of an outbreak signal into three series at the same time (either all three GI series or all three Resp series).

GI

	uniHW	uniReg	multiReg	mahalanobisHW	normsumHW	PCAHW
at28	0.040	0.047	0.045	0.053	0.047	0.034
at56	0.020	0.020	0.025	0.023	0.016	0.014
at72	0.011	0.019	0.017	0.019	0.012	0.008
in14_112	0.039	0.047	0.049	0.050	0.044	0.037

Resp

	uniHW	uniReg	multiReg	mahalanobisHW	normsumHW	PCAHW
at28	0.031	0.034	0.043	0.043	0.039	0.039
at56	0.012	0.019	0.017	0.022	0.022	0.019
at72	0.011	0.016	0.016	0.016	0.022	0.014
in14_112	0.036	0.040	0.044	0.046	0.047	0.041

Table 4-3: All Series Outbreak Detection Rates

Detection performance on an outbreak inserted into all series of a particular syndrome type. Recall that the metrics for evaluation are Detection Rate at a fixed false alarm rate of once every 28 days; Detection Rate at a fixed false alarm rate of once every 56 days; Detection Rate at a fixed false alarm rate of once every 72 days; and integrated Detection Rate (area under the ROC curve) for false alarm rates between every 14 to 112 days. Normsum refers to the Standardized Mean described in 4.2.1.1. The best performance in each row is shaded.

A few observations are clear from the results in Table 4-3. First, attempting to detect purely from a single normalization on a single series is not very effective compared to other methods. Second, multivariate regression provides some improvement when the series it uses are related, even when the outbreak occurs only in one of those series. Third, the Mahalanobis combination seems to be a distinct improvement for detecting multivariate outbreaks.

4.2.3. Conclusions and Future Work

Using multivariate data streams is clearly valuable tool improving detection. The experiments here show how using multiple data streams can provide this improvement, especially when the outbreak may cause a signal in more than one of them. However, little is currently known about the appearance of different diseases in specific health data series; in particular, given the fact that the outbreak signal is

likely to be different in different series, the study could be expanded to consider different patterns or ranges of patterns. In addition, given that we know the size of the outbreak will change the performance difference between algorithms, studying a broader range of outbreak signal sizes would provide understanding of the impact of multivariate methods for different sizes. As this is investigated further, the methods presented here can be tuned specifically for distinct types of diseases and health series.

Other combination and assistance methods can also be considered, such as an ARIMA model as a predictive method (such that lagged factors might be considered) or the use of burst detection methods from text analysis (Kleinberg, 2003). In these results, PCA performance was shown to be relatively poor; however, only the first principal component was used, and so a PCA method which uses more components might be more effective. Another direction of multivariate work would be to directly monitor the covariances in a sliding covariance window, in a fashion similar to the moving-F test (Riffenburgh & Cummins, 2006). In addition, a comparison method of alerting if any of the univariate series alerts might be useful for comparing outbreaks which occur across multiple series. Finally, these methods should be directly compared to the MCUSUM and MEWMA methods, as well as the directional T^2 method, on a larger variety of outbreak types (especially smaller outbreaks) in order to provide a better understanding of their relative performances.

4.3. Additional Day-of-week Preprocessing for Detection Improvement

4.3.1. Method Description

In Chapter 2, we discussed the relationship between forecasting accuracy and detection performance. However, we also discussed a number of forecast residual attributes which can negatively impact the detection performance. One of these was seasonal variance, particularly day-of-week seasonal variance. In this case, the variance of the residuals differs by day-of-week (usually with lower variance on weekends). One way to come closer to the standard iid normal paradigm is to scale all residuals by a day-of-week factor, in order to have a common variance. This does not fit into the forecasting paradigm in Chapter 2, but is a post-forecasting method for improving detection.

Here, we investigate a method for rescaling residuals by day-of-week variance. Specifically, we consider estimating the variance of each day-of-week using residuals from the past. We assume that the forecasting process creates residuals with seven different standard deviations, such that $r_t \sim N(0, \sigma_i)$, *i=t mod 7 +1*. Then, using the past residuals, create estimates $\hat{\sigma}_1 \ldots \hat{\sigma}_7$. Finally, for the current day, instead of using the residual r_t, use the scaled residual $r_t / \hat{\sigma}_i$. The idea is to (approximately) standardize the variance of each day, such that residuals on different days will have equal variance (approximately equal to 1). This should improve the detection by reducing the day-of-week seasonal variance described in Chapter 2.

We also consider using only positive residuals. The motivation for using only the positive aberrations has two reasons. First, it gives a better estimate of the variance in the aberrations to be detected: those where the count is higher than expected; due to the imperfect nature of the forecaster, this can be different for over-predictions versus under-predictions. Second, it implicitly avoids including negative singularities, points where the actual count is much lower than expected, often zero or nearly zero (due to holidays or other factors); when such negative singularities are included, they significantly increase the estimated standard deviation (thus reducing the scaling of any outbreak occurring on those days-of-week, usually Mondays).

4.3.2. Empirical Test Results

In order to determine whether or not this day-of-week preprocessing method is useful, we want to see if it improves outbreak detection. To do this, we compare Detection Rates over a variety of false alert levels, either with or without the various day-of-week variance preprocessing. By doing this, we can measure the impact of using the preprocessing technique.

To evaluate the effectiveness of this technique, we take the BioALIRT data set described in Section 1.3.1 and examine each of the six series. For each series, we use a Holt-Winters forecaster (described in 3.2.3) to generate residuals. Next, we compare three post-forecasting methods:

Standard: Using the unchanged residuals (standard Holt-Winters);

DOW-SD: Using all past days to estimate day-of-week seasonal variances and dividing by the estimated standard deviation for the current day;

Positive DOW-SD: Using only past days with *positive* residuals to estimate the day-of-week seasonal variances and dividing by the estimated standard deviation for the current day.

We then use a Shewhart chart to monitor each series and to generate alerts.

We generate lognormal outbreak signals and insert them into each possible day in the data series (after day 250, in order to allow sufficient data to estimate the day-of-week standard deviations and positive standard deviations). Table 4-4 shows the Detection Rates for the three variants on the six series, for a variety of outbreak sizes and false alert rates. Figure 4-1 through Figure 4-3 show selected ROC curves for different series and outbreak sizes.

Average Detection Rate		method		
cases	FA rate	standard	dowSD	positiveDOWSD
20	1/112	0.04	0.04	0.04
	1/56	0.08	0.08	0.07
	1/28	0.15	0.15	0.16
	1/14	0.3	0.29	0.28
20 Total		0.14	0.14	0.14
50	1/112	0.04	0.04	0.05
	1/56	0.08	0.09	0.09
	1/28	0.14	0.16	0.18
	1/14	0.29	0.3	0.32
50 Total		0.14	0.15	0.16
100	1/112	0.07	0.09	0.12
	1/56	0.12	0.15	0.18
	1/28	0.19	0.25	0.28
	1/14	0.36	0.4	0.44
100 Total		0.18	0.22	0.25
200	1/112	0.18	0.22	0.28
	1/56	0.24	0.31	0.34
	1/28	0.34	0.42	0.44
	1/14	0.53	0.59	0.62
200 Total		0.32	0.38	0.42
400	1/112	0.35	0.38	0.46
	1/56	0.38	0.52	0.5
	1/28	0.51	0.67	0.64
	1/14	0.71	0.82	0.8
400 Total		0.48	0.6	0.6

Table 4-4: Day-of-Week Normalization Detection Rates

Average detection probabilities for each method, over the six BioALIRT data series, for a variety of false alert rates and outbreak sizes. Each entry is the average detection probability, over all six series, for a particular method, on a particular outbreak size and false alert rate. The totals give the average probability of detection over all false alert rates for that outbreak size. For example, monitoring standard Holt-Winters residuals, in an outbreak size of 400 and false alerts every 14 days (FA Rate= 1/14), has a Detection Rate of 0.71 (detects 71% of outbreaks) . Across all four false alert rates (1/112, 1/56, 1/28, and 1/14), for outbreaks of size 400, it has an average Detection Rate of 0.48. If the process including normalizing by day-of-week standard deviation is used, the Detection Rates are 0.82 and 0.60, respectively.

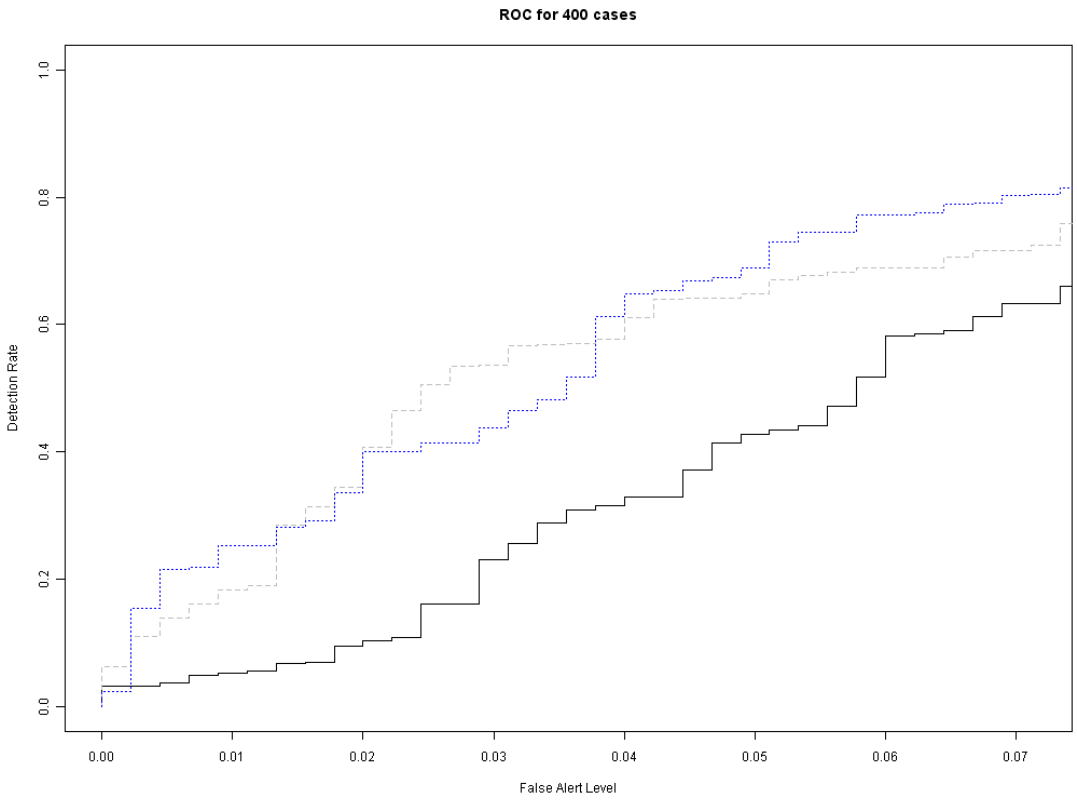

Figure 4-1: ROC Curves for Day-of-week Residual Normalization on Resp/400
ROC curves for the standard HW residuals (solid black), residuals normalized by day-of-week standard deviation (dashed grey), and residuals normalized by positive day-of-week (dotted blue). This figure shows results for Respiratory Military Visits, with an outbreak of total size 400.

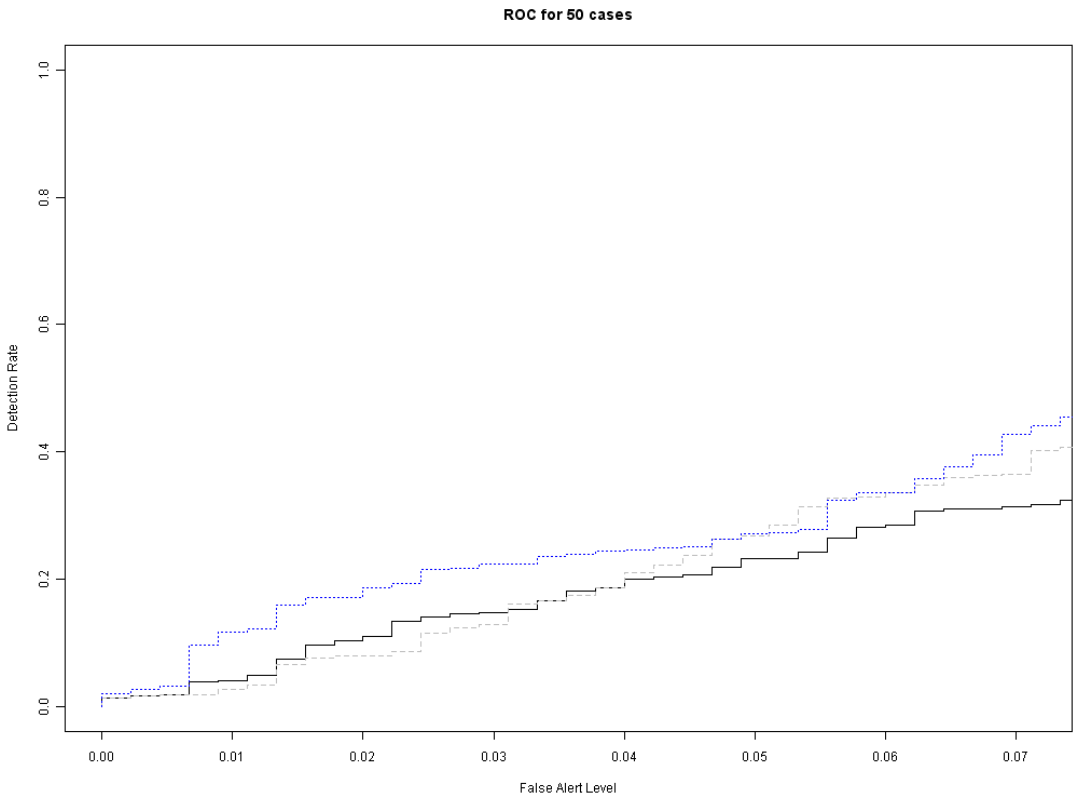

Figure 4-2: ROC Curves for Day-of-week Residual Normalization on GI/50
ROC curves for the standard HW residuals (solid black), residuals normalized by day-of-week standard deviation (dashed grey), and residuals normalized by positive day-of-week (dotted blue). This figure shows results for GI Civilian Visits, with an outbreak of total size 50.

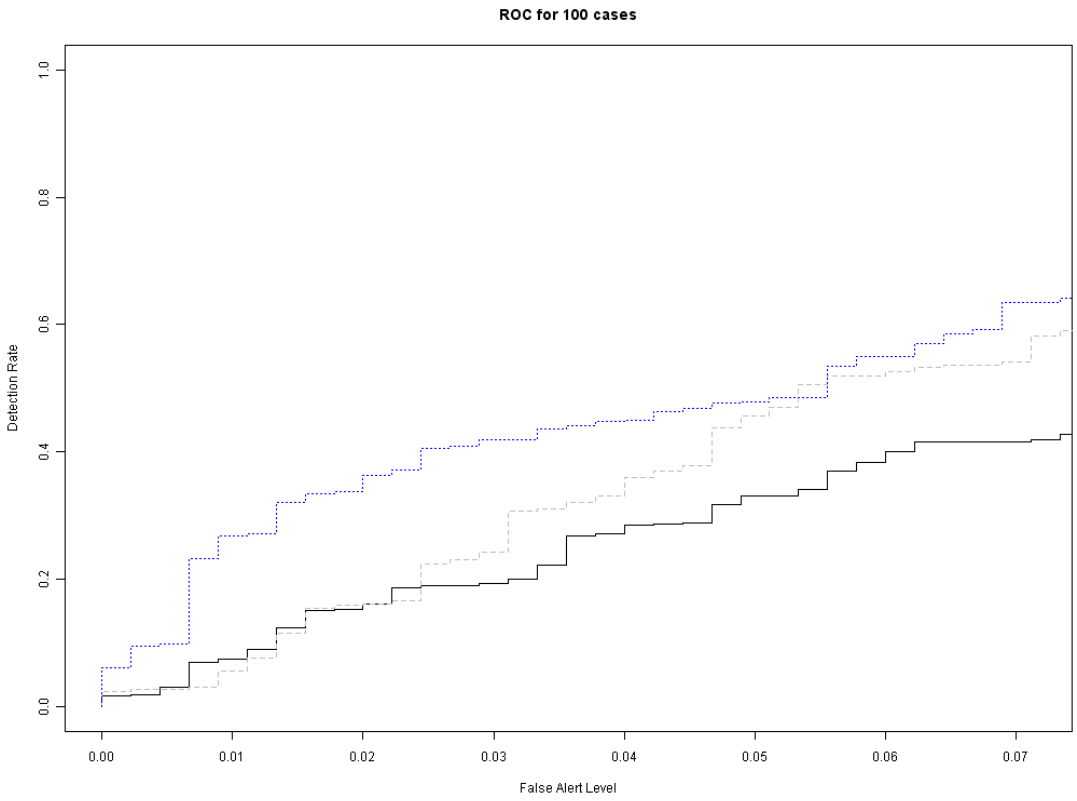

Figure 4-3: ROC Curves for Day-of-week Residual Normalization on GI/100
ROC curves for the standard HW residuals (solid black), residuals normalized by day-of-week standard deviation (dashed grey), and residuals normalized by positive day-of-week (dotted blue). This figure shows results for GI Civilian Visits, with an outbreak of total size 100.

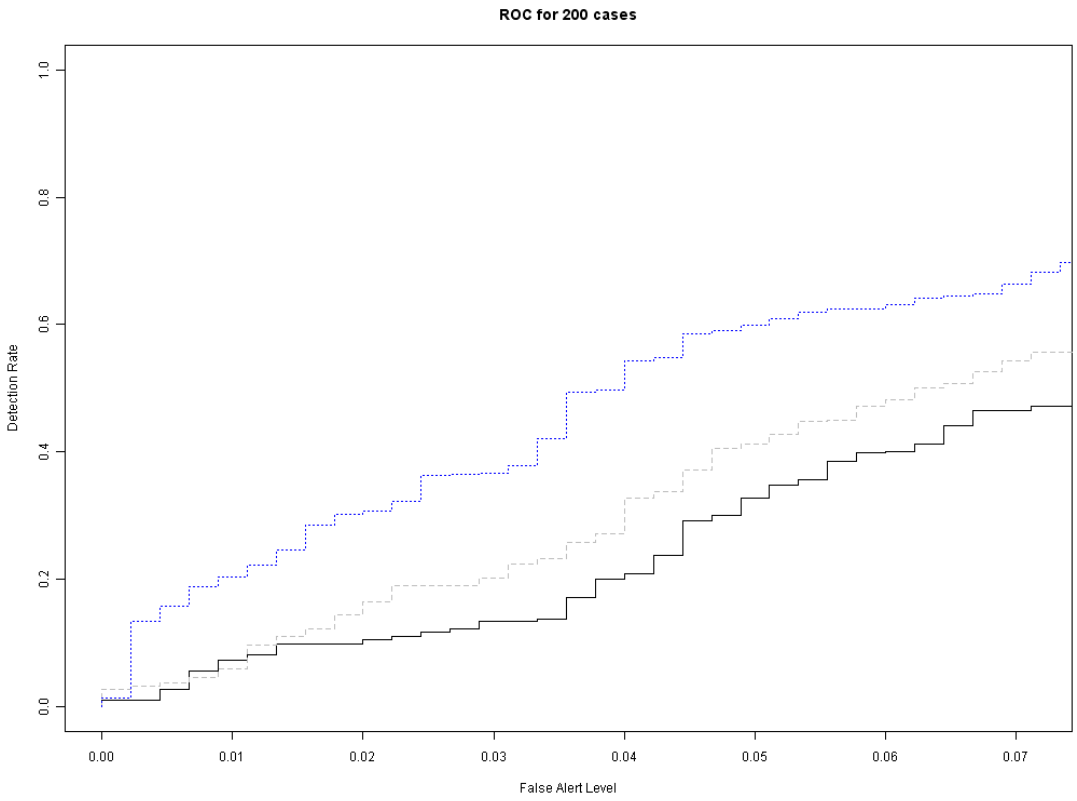

Figure 4-4: ROC Curves for Day-of-week Residual Normalization on GI/200
ROC curves for the standard HW residuals (solid black), residuals normalized by day-of-week standard deviation (dashed grey), and residuals normalized by positive day-of-week (dotted blue). This figure shows results for Gastrointestinal Prescriptions, with an outbreak of total size 200.

From the figures and table, we can see that there is substantial improvement from using the day-of-week standardization; this is particularly pronounced for low false alert rates and large outbreak sizes. This seems to be due to the fact that on low-variance days, the outbreak can be much more clearly seen as an aberration, and thus much more easily detected.

This effect can be seen in Figure 4-5 and Figure 4-6, which compare a normal Holt-Winters/Shewhart detection without any post-processing, versus using a positive-value day-of-week standard deviation normalization. The difference in effect can be

seen to be largely due to weekday/weekend difference. We can see that the weekdays are significantly improved, especially at the low false alert levels. While the Detection Rates actually decrease on weekdays, this is more than made up for by the increase in weekend detection

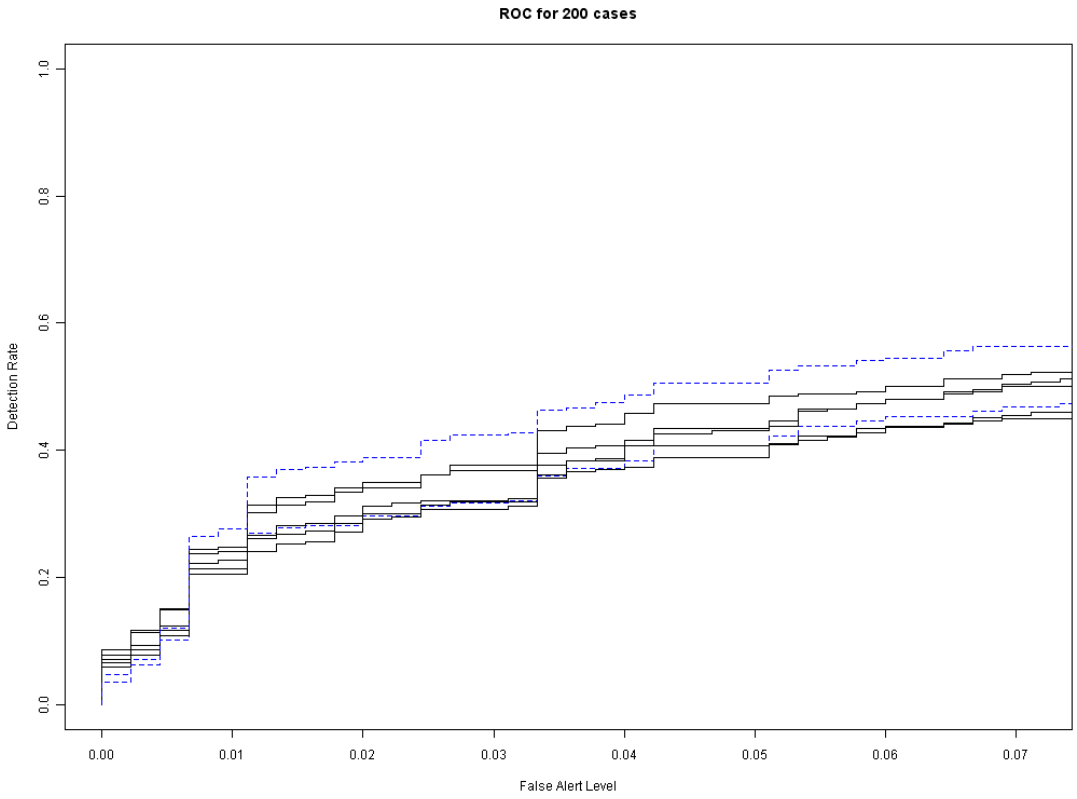

Figure 4-5: ROC by Day-of-week for Holt-Winters
ROC for Gastrointestinal Military Visits, using a standard Holt-Winters forecast with Shewhart detection. Results are displayed separately by day of week: weekdays are solid black lines, weekends in dashed blue lines. We can see that each day of week is approximately the same rate of detection, depending mainly on the size of the outbreak and the false alert rate, rather than the size of the outbreak relative to the normal size for the day of week.

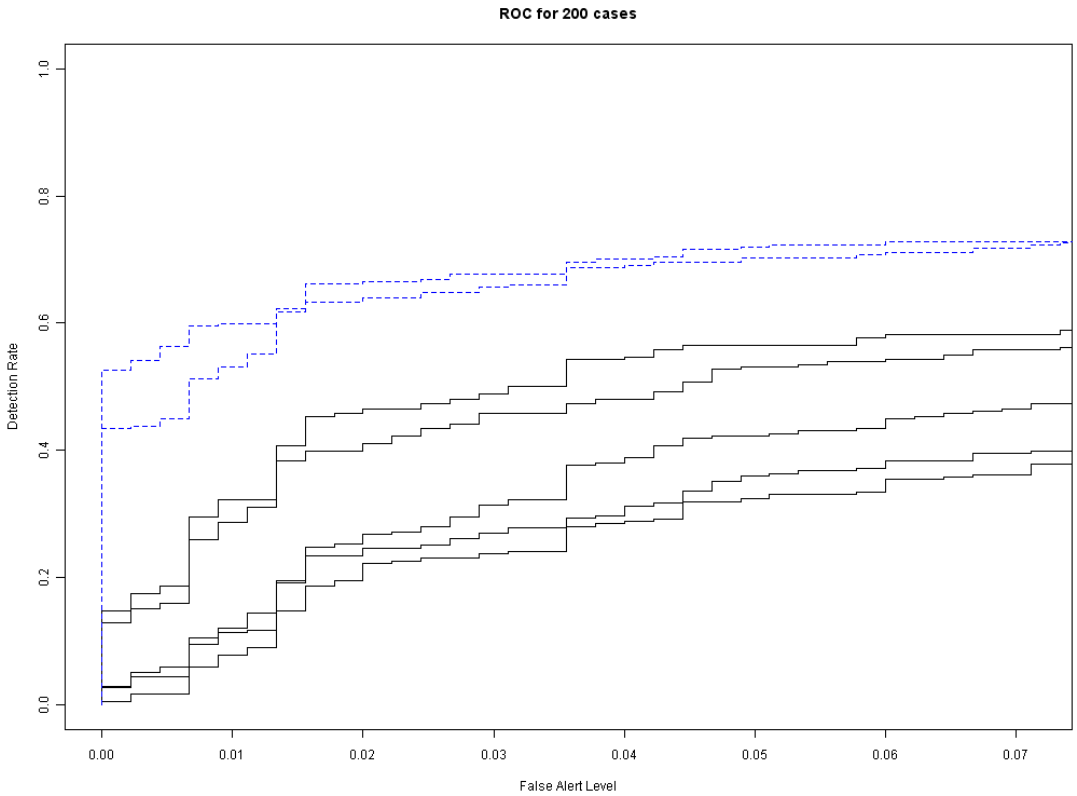

Figure 4-6: ROC by Day-of-week for Holt-Winters with Day-of-week Residual Normalization ROC for Gastrointestinal Military Visits, using a standard Holt-Winters forecast with Shewhart detection, followed by a day-of-week positive value standard deviation normalization. Results are displayed separately by day of week: weekdays are solid black lines, weekends in dashed blue lines.

Using the variance of the positive aberrations is at least as good as using all residuals (assuming sufficient data are available), and sometimes gives significant improvement.

4.3.3. Conclusions and Future Work

It seems that scaling by the estimated standard deviation of the positive residuals is an improvement over scaling by the estimated standard deviation of all residuals from the same day. This technique depends on having enough data to estimate the day-of-week standard deviations, but once this is available, it seems to result in a significant and unambiguous improvement for detection.

It is not clear from these results how much of the improvement is due to a closer fit to the population of interest (days with positive residuals) and how much is due to removing negative singularities (outliers with very low negative residuals). Further investigation could reveal this, and suggest the usefulness of a more robust method of estimating the standard deviation (such as the median absolute deviation). But we strongly suggest adding this method to existing detection algorithms, as it is likely to provide simple but marked improvement in Detection Rates.

4.4. Efficient Detectors

4.4.1. Efficient Scores and The CuScore Method

We now show how we can use efficient statistics and the CuScore method to find an improved method for detecting outbreaks. The idea of the efficient statistic is due to Fisher (Fisher, 1922), who recognized that any unbiased statistic can be evaluated in terms of how well its variance approaches the Cramer-Rao lower bound. Since any statistic $\hat{\theta}$ estimating an unknown parameter θ has a certain variance $\sigma^2_{\hat{\theta}}$, and the lowest variance of any unbiased estimator can be bounded below by the Cramer-Rao lower bound, $1/I(\theta)$, the *efficiency* of a statistic can be measured as the ratio of these two factors, $e(\hat{\theta}) = \frac{1/I(\theta)}{\sigma^2_{\hat{\theta}}}$, with an *efficient* statistic being one which achieves the lower bound and thus an efficiency of 1.

A score is a number generated for a timeseries, which is then checked to determine if the timeseries is in control. *Efficient* scores are simply statistics which are efficient

for testing a null hypothesis versus an alternative. This is particularly relevant to anomaly detection, as we are attempting to test when an anomaly has occurred (the alternative) versus normal background variability (the null hypothesis). Thus, in process monitoring, the CuScore is a method for determining the efficient statistic testing whether a process has gone out of control in a specific way (Box & Luceno, 1997). A CuScore statistic can be constructed for testing any fixed deviation from the standard normal white noise assumptions, and standard control charts are efficient score statistics for various kinds of deviations. The Shewhart chart is optimized for detecting a single-day spike outbreak. The CuSum efficiently detects a continuing step increase. The EWMA detects an exponential increase. Finally, a moving average of the last k days efficiently detects a temporary step increase lasting k days. These can be seen in Figure 4-7.

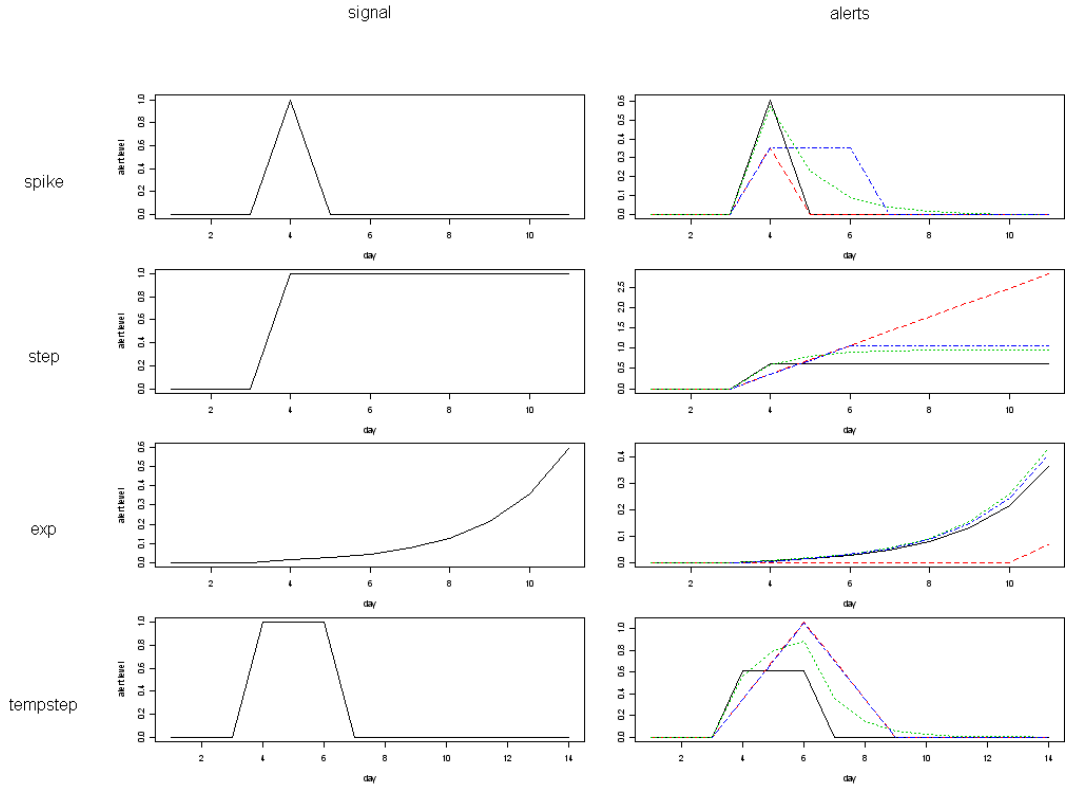

Figure 4-7: Daily Scores for Various Detection Methods
The scores resulting from using various detection methods on various signals, shown without noise. Shewhart is black, EWMA (with $\lambda = 0.6$) is dotted green, CuSum is dashed red, and a 3-day moving average is dot-dashed blue. In order to indicate the strength of different detection methods on different signals, all scores are normalized to have a threshold of 1 and a false alert level of 1/20 under standard normal data.

The CuScore is a method for determining the efficient score for a known type of signal in a time series of white noise. Its effectiveness lies in the fact that when the signal of interest occurs, the residuals will contain a component which correlates with the CuScore detector. Described in (Box & Ramirez, 1992), it proceeds as follows:

1. Formulate the null model as $y_t = \hat{y}_t + a_{t0}$, where $\hat{y}_t$ is the estimated or forecasted value for y_t and a_{t0} is white noise error.

2. Define the signal of interest, $o = [[o_i]], i = 1...I$ and form the discrepancy model, $y_t = \hat{y}_t + \delta o_{t-I+i} + a_t$, where a_t is white noise error.

3. Compute the CuScore as $Q_t = \sum_{i=1}^{I} (y_{t-I+i} - \hat{y}_{t-I+i}) o_i$.

4.4.2. CuScore for a Lognormal Outbreak

A lognormal progression is a reasonable model for an outbreak signal, because as (Burkom, 2003a) describes, the incubation period distribution of many infectious diseases can be approximated well by a lognormal distribution, with parameters dependent on the disease agent and route of infection. Thus, it is reasonable to assume that outbreaks of such diseases will result in the addition of a lognormal number of cases to the normal background cases in the health series. In this case, we take as our target signal a lognormal outbreak with shape parameter σ and scale parameter m (note that this can also be reparameterized using $\mu = \ln(m)$ as the mean of the log, and σ as the standard deviation of the log). The lognormal density is then multiplied by k, the total number of infected cases, to give a distribution of the total number of people expected to be symptomatic at each point. This multiplied curve is then binned by day and rounded, to provide a daily count for the number of people who would be symptomatic and added to the observed health series daily count. For an illustration of the process, see Figure 4-8.

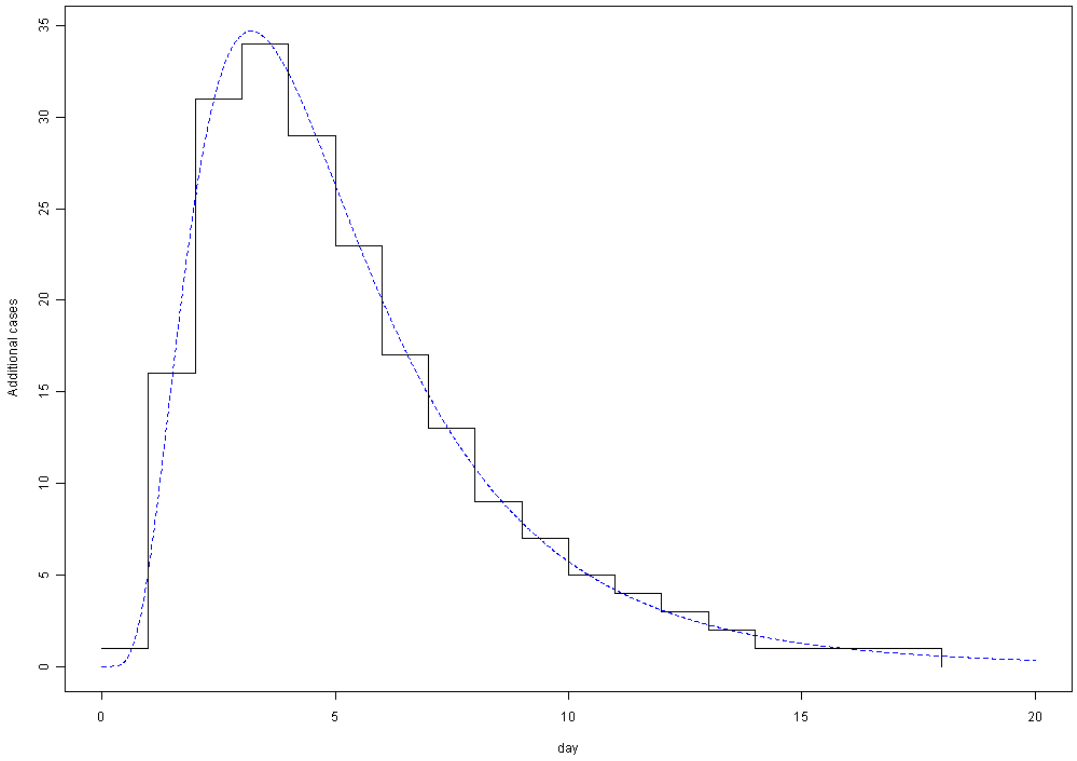

Figure 4-8: Binned Lognormal Outbreak
Lognormal outbreak (in dashed blue), binned into daily additional counts. The maximum occurs on day 4.

This use of the lognormal distribution to model disease outbreaks is known in the biosurveillance literature, and many recent biosurveillance evaluations have used lognormal curves to approximate the disease outbreak signal (Burkom, 2003b, Burkom et al., 2007). However, none have attempted to directly build an optimal detector for lognormal outbreaks. Here, we use the CuScore method to build such an optimal detector.

In the case of early detection, it is only relevant to detect up to the point of maximum infection; intervention up to this point can have a significant impact on the public health effect, but later intervention has minimal effect, as the infected population is

already naturally recovering. Thus, we only consider the days up to the day containing the mode for the outbreak signal lognormal distribution, which for a lognormal can be found at $e^{\mu-\sigma^2}$. Our CuScore detector will therefore be a weighted sum of the past D days' residuals, where $D = \lceil e^{\mu-\sigma^2} \rceil$ and each residual $r_t = y_t - \hat{y}_t = y_t - f_t$.

$$c_t = \sum_{d=1}^{D} \left(\int_{d-1}^{d} \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln(x)-\mu)^2}{2\sigma^2} \right] dx \right) r_{t-D+d} \quad \text{(Eq. 4-6)}$$

This set of weights will have the maximum correlation with a binned lognormal signal over the past D days, as it uses the expected values for the lognormal signal as daily weights. Under normality assumptions, its variance can be calculated as a linear combination of normal variables, $V(c_t) = w^T \Sigma w$, where Σ is the covariance matrix of D days of residuals ($\sigma^2 I$ if they are independent and more complex if there is autocorrelation) and w is the weight vector,

$w = [[w_i]], w_i = \int_{i-1}^{i} \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln(x)-\mu)^2}{2\sigma^2} \right] dx.$

4.4.3. Optimizing CuScore for Timeliness

However, the CuScore alone is not an optimal detector for a signal. We can see this in Figure 4-9, which compares the ROC curves for a Shewhart, CuSum, and Lognormal CuScore detector for a Lognormal outbreak.

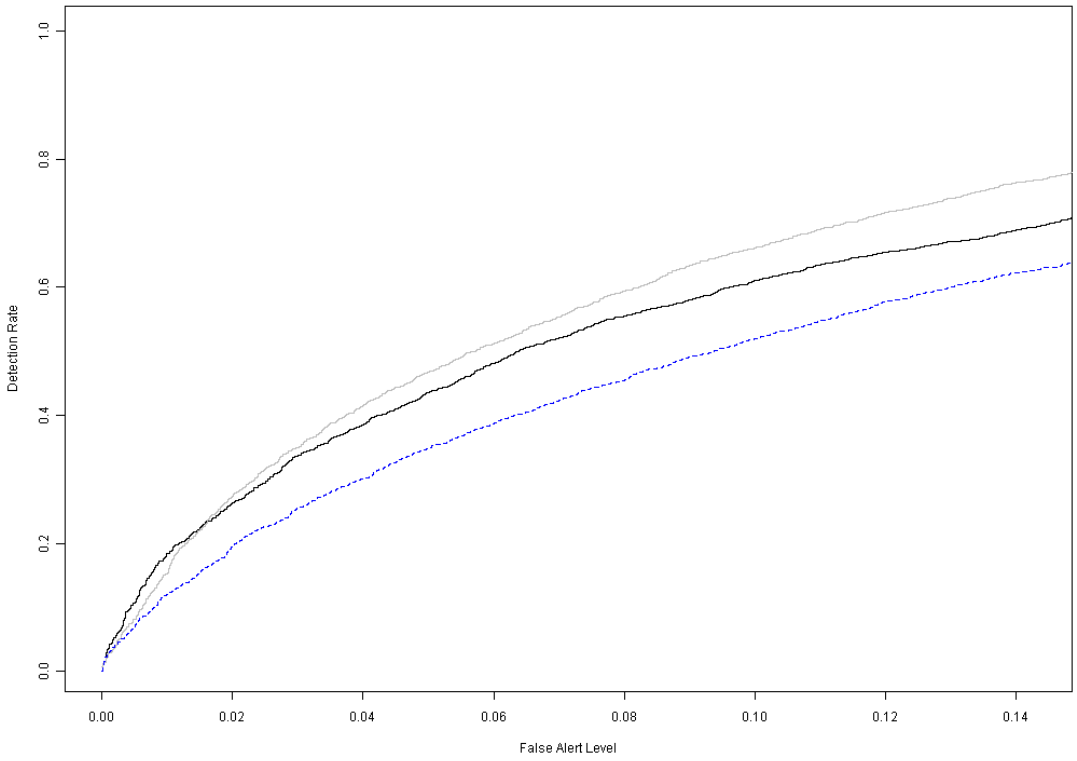

Figure 4-9: ROC for CuScore, Shewhart, and CuSum on Lognormal Outbreak
ROC curves for CuScore (in black), Shewhart (in grey), and CuSum (in dashed blue), for a lognormal outbreak of total size 200 (peak size of 34). Only the portion up to false alert rate 1/14 is shown.

Although it is an efficient detector, the CuScore detector is unexpectedly dominated by the Shewhart. This is because it does not address the issue of detection over time. As can be seen in the earlier signals (in Figure 4-7) and in Figure 4-10, the CuScore method maximizes the score at the *end* of the signal it is optimized for. We can see the Shewhart maximized at the spike, CuSum continues to increase after the step, and the lognormal CuScore maximizes the score at the *end* of the lognormal.

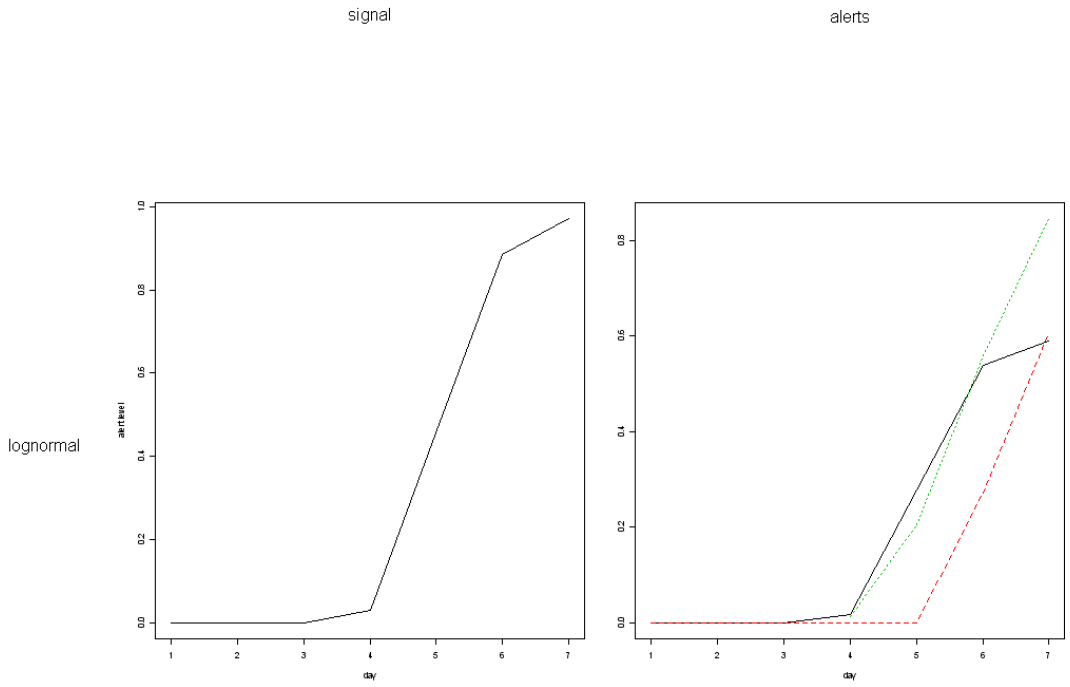

Figure 4-10: Daily Scores on Lognormal Outbreak
A lognormal signal and the corresponding scores for detection for a Shewhart (solid black), CuSum (dashed red), and CuScore (dotted green).

For a fixed-length outbreak, the efficient score detector is optimizing for a detection on the *last* day of the observed signal. We can see this result in Figure 4-11, which compares the ROC curves, using *only* the fourth day (the day of the outbreak peak, the end of our detection curve) to detect.

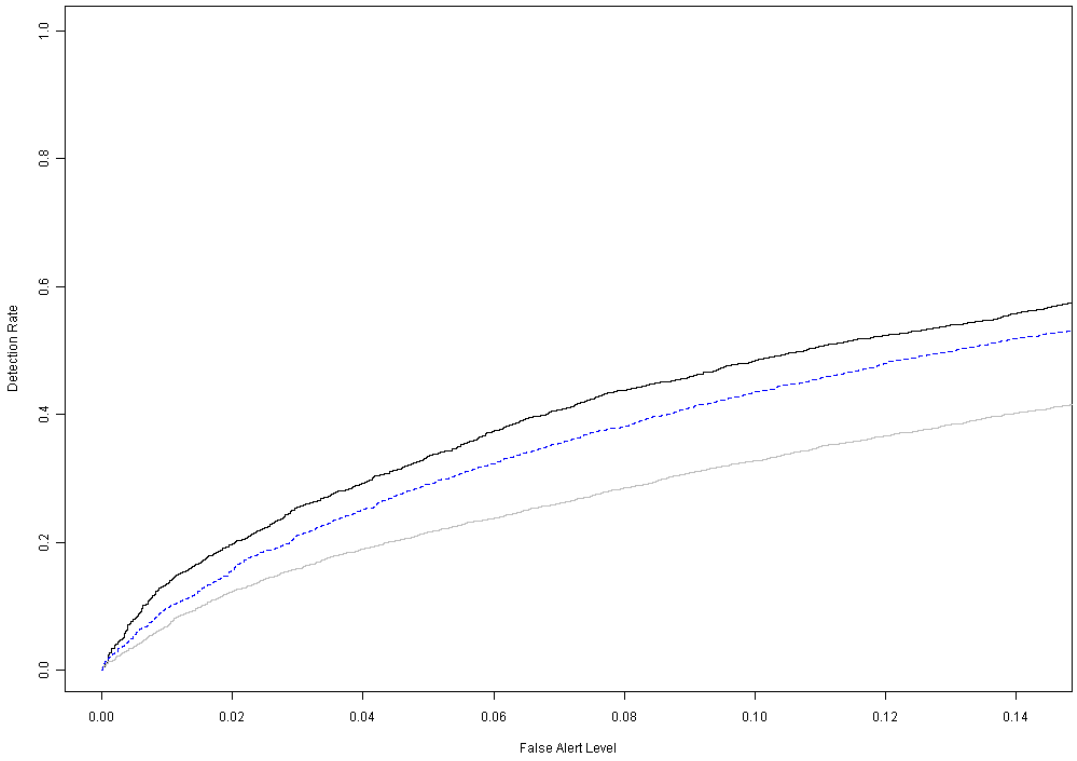

Figure 4-11: Day-4 Only ROC for CuScore, Shewhart, and CuSum on Lognormal Outbreak ROC curves for CuScore (in black), Shewhart (in grey), and CuSum (in dashed blue), for a lognormal outbreak of total size 200 (peak size of 34). Only the portion up to false alert rate 1/14 is shown. In this case, only the fourth day is used for detection, at the peak of the outbreak. When only the fourth day is considered, the CuScore is indeed the most effective at detecting the outbreak.

While in a single-day comparison on the final day, the CuScore method will be efficient, this is not the most effective way of detecting an outbreak. The CuScore analysis fails to account for the fact that each day is a separate test, and so there are multiple chances to detect the outbreak and (ideally) detect it even earlier. If we only detect at the end of the outbreak, then we have failed to provide timely warning so that action can be taken. Figure 4-9, showing the ROC curves for a CuScore method versus Shewhart and CuSum, illustrates this. Even though the CuScore method is designed to detect this specific type of outbreak, because it is only optimized for the

last day of detection, it reduces the chances to detect it before the last day, and thus results in a lower Detection Rate than Shewhart over the course of several days.

An optimization approach is thus suggested: one wishes to minimize the average day of detection, for a given false alert rate, given a specific expected outbreak signal. For any type of outbreak, we can consider the maximum useful day of detection as the last day at which action will still be useful, or the first day at which lab reports will confirm the disease. This can then be used as the maximum delay, for determining the cost of not alerting earlier (effectively missing the disease outbreak). This can be formulated as an optimization problem, where the objective function is to minimize the expected time of detection. The expected time of detection can be computed as the weighted sum of delays, using the probability of detection on each day as weight, using $D + 1$ when the method fails to detect the outbreak:

$$\sum_{d=1}^{D} dp'_d + (D+1)(1 - \sum_{d=1}^{D} p'_d) \qquad \text{(Eq. 4-7)}$$

In this equation, p'_d is the probability of detection on day d, conditional on not having detected on an earlier day.

Alternatively, one may wish to maximize the overall Detection Rate in the first D days. This can be done by using a slightly modified function to optimize, $\sum_{d=1}^{D} p'_d$.

4.4.4. Direct Solutions using the Multivariate Normal Distribution

We now consider again, under this optimization framework, the class of weighted sums of (residual) observations from the past D days. We wish to optimize over this class, resulting in a set of weights. The weighted sum over the past D days will then

be a detector with the smallest mean time of detection (for a given outbreak shape and false alert rate).

In order to find a closed form solution for optimizing the detector, we make the following simplifying assumptions: that the daily (residual) baseline health value is an iid normal variable, with mean 0 and common standard deviation σ. We can then use the fact that each day's alert value comes from a multivariate normal distribution. If $w = [w_1 ... w_D]$ is the set of weights over the past D days, and $r = [r_1 ... r_t]$ are the set of residual values from the health series, then the alert values over the past D days will be the application of the weights over the sliding window of daily values, $c = [c_{t-D+1} ... c_t] = [w^T(r_{t-2(D-1)} ... r_{t-D+1}) ... w^T(r_{t-D+1} ... r_t)]$. Since

$c_t = \sum_{d=1}^{D} w_d r_{t-D+d}$, if the underlying health data residuals are approximately normal, with common variance, then each c_t is normally distributed, and collectively both c and y have a multivariate normal distribution. More specifically, the means and variances will depend on whether or not there is an outbreak. If there is no outbreak, then. $\begin{pmatrix} (r_{t-2(D-1)} ... r_t^T \\ c^T \end{pmatrix}$ will have 0 mean and covariance matrix

$\Sigma = \sigma^2 \begin{pmatrix} I & A \\ A & C \end{pmatrix}$, where, assuming no correlation in r_i, I is the identity matrix $I_{(2D-1,2D-1)}$,

$$A = \begin{pmatrix} w_D & 0 & 0 & ... & 0 & ... & 0 & 0 \\ w_{D-1} & w_D & 0 & ... & 0 & ... & 0 & 0 \\ w_{D-2} & w_{D-1} & w_D & ... & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & w_1 & w_2 & ... & w_D & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & w_1 & ... & w_{D-1} & w_D \end{pmatrix}_{(D,2D-1)}$$

and

$$C = \begin{pmatrix} \sum_{i=1}^{D} w_i w_{i-0} & \sum_{i=2}^{D} w_i w_{i-1} & \sum_{i=3}^{D} w_i w_{i-2} & ... & \sum_{i=D-1}^{D} w_i w_{i-(D-2)} & w_D w_1 \\ \sum_{i=2}^{D} w_i w_{i-1} & \sum_{i=1}^{D} w_i w_{i-0} & \sum_{i=2}^{D} w_i w_{i-1} & ... & \sum_{i=D-2}^{D} w_i w_{i-(D-3)} & \sum_{i=D-1}^{D} w_i w_{i-(D-2)} \\ ... & ... & ... & ... & ... & ... \\ w_D w_1 & \sum_{i=D-1}^{D} w_i w_{i-(D-2)} & \sum_{i=D-2}^{D} w_i w_{i-(D-3)} & ... & \sum_{i=2}^{D} w_i w_{i-1} & \sum_{i=1}^{D} w_i w_{i-0} \end{pmatrix}_{(D,D)}$$

Using this, we could determine the probabilities of various run lengths under the no-outbreak state. For the most part, since we are not using runs rules, we will only be concerned with the individual-day variance ($V(c_t) = \sigma_c^2 = \sigma^2 \sum_{i=1}^{D} w_i^2$) to control the FA rate.

However, if there is an outbreak during the past D days, then the distribution changes due to the additional outbreak cases. In particular, let the outbreak $o = (o_1, o_2 ... o_D)$. The covariance matrix remains the same, but the mean of c is now $E(c) = (w_D o_1, \sum_{i=1}^{2} w_{D-i+1} o_i, ..., \sum_{i=1}^{D} w_{D-i+1} o_i)$. Given this, we can determine the probability of at least one day's alert value c_i being sufficient to alert, by considering only the marginal distribution of c, which will also be multivariate normal, $c \sim N(E(c), C)$. While the cumulative distribution function for the multivariate normal distribution is intractable to solve for exactly, it can be numerically calculated. We use R's mvtnorm package (Genz et al., 2009) to determine the probability that at least one of the days provides an alert during the outbreak and also to find the expected probability of detection on each day of the outbreak (thus providing the mean delay before detection).

To determine the optimal detector for a given outbreak signal and false alert level, the optimization problem is to maximize the probability that at least one day's weighted detection value is high enough to alert. More formally: given false alert rate α, residual variance σ^2, and outbreak shape $o = (o_1, o_2, ..., o_D)$,

maximize $1 - \Phi_D(UCL(w), EC(w), C(w))$

s.t. $\sum_{i=1}^{D} w_i = 1$ and

$1 - \Phi(UCL(w)/SD(w)) \leq \alpha$

where $UCL(w) = SD(w)\Phi^{-1}(1-\alpha)$,

$$C(w) = \begin{pmatrix} \sum_{i=1}^{D} w_i w_{i-0} & \sum_{i=2}^{D} w_i w_{i-1} & \sum_{i=3}^{D} w_i w_{i-2} & \cdots & \sum_{i=D-1}^{D} w_i w_{i-(D-2)} & w_D w_1 \\ \sum_{i=2}^{D} w_i w_{i-1} & \sum_{i=1}^{D} w_i w_{i-0} & \sum_{i=2}^{D} w_i w_{i-1} & \cdots & \sum_{i=D-2}^{D} w_i w_{i-(D-3)} & \sum_{i=D-1}^{D} w_i w_{i-(D-2)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ w_D w_1 & \sum_{i=D-1}^{D} w_i w_{i-(D-2)} & \sum_{i=D-2}^{D} w_i w_{i-(D-3)} & \cdots & \sum_{i=2}^{D} w_i w_{i-1} & \sum_{i=1}^{D} w_i w_{i-0} \end{pmatrix}_{(D,D)},$$

$SD(w) = \sqrt{V(c_t)} = \sqrt{\sigma^2 \sum_{i=1}^{D} w_i^2}$, and

$EC(w) = (E(c_1), E(c_1), ..., E(c_D)) = (w_D o_1, \sum_{i=1}^{2} w_{D-i+1} o_i, ..., \sum_{i=1}^{D} w_{D-i+1} o_i)$.

Φ_D is the cumulative density function for the multivariate normal distribution,

$\Phi_D(UCL, \vec{\mu}, C) = \frac{1}{\sqrt{|C|(2\pi)^D}} \int_{-\infty}^{UCL} \int_{-\infty}^{UCL} ... \int_{-\infty}^{UCL} exp(-\frac{1}{2}(\theta - \mu)^T C^{-1}(\theta - \mu) d\theta$.

4.4.5. An Optimized Lognormal CuScore

As an example of this method, we present a detection ensemble which is optimized to provide the maximum probability of detection for a lognormal outbreak. In this case, as in Section 4.4.2, we take as our target signal a lognormal outbreak with shape parameter σ and scale parameter m, binned and rounded appropriately. Since the objective function to be minimized is nonlinear, we use the limited-memory BFGS quasi-Newton method described by (Byrd et al., 1995), constraining each weight to

be between 0 and 1. We then normalize to have all weights sum to 1. This was done for several false alert levels and several sizes of lognormal outbreak. Some weightings determined by this approximation-maximization method are shown in Table 4-5.

FA Rate	ncases	weight1	weight2	weight3	weight4	Improvement vs.Shewhart
1/14	10	0.00	0.00	0.00	1.00	
1/14	100	0.00	0.00	0.02	0.98	0.02%
1/14	1000	0.00	0.00	0.00	1.00	
1/14	10000	0.00	0.00	0.00	1.00	
1/28	10	0.00	0.00	0.00	1.00	
1/28	100	0.00	0.00	0.17	0.83	1.24%
1/28	1000	0.00	0.15	0.37	0.48	
1/28	10000	0.00	0.00	0.00	1.00	
1/56	10	0.00	0.00	0.00	1.00	
1/56	100	0.00	0.05	0.18	0.76	3.76%
1/56	1000	0.00	0.15	0.37	0.48	0.03%
1/56	10000	0.00	0.00	0.00	1.00	
1/112	10	0.00	0.00	0.00	1.00	
1/112	100	0.01	0.10	0.21	0.68	6.96%
1/112	1000	0.00	0.17	0.38	0.45	0.11%
1/112	10000	0.00	0.00	0.00	1.00	

Table 4-5: Optimal Detection Weightings for Lognormal
The optimal weightings (over the past four days) found for optimizing detection of a lognormal outbreak which peaks on the fourth day. The number of total additional cases due to the outbreak is indicated in the ncases column. Rows which improve over Shewhart are highlighted.

Interestingly, the optimal weighting combination depends on both the size of the lognormal outbreak and the false alert level allowed. For large outbreaks or a high false alert rate, a Shewhart actually becomes most timely, as it has a reasonable chance of detection on several days. But for smaller outbreaks of the same shape, or in situations requiring a lower false alert level, the optimal weights tend towards the lognormal CuScore weights. In order to have a chance of detecting a small outbreak, with low false alert rate, greater sensitivity to the outbreak is needed. This is best

achieved by maximizing the Detection Rate on the last day of the outbreak (when there is full information).

The shape of the outbreak also plays a role; for outbreaks which peak later, the optimization method can result in an even more dramatic improvement over the standard Shewhart detection. For example, for a lognormal outbreak which peaks 14 days after starting, improvement can be nearly 50% in some cases. Table 4-6 shows detection results for a longer outbreak.

FA Rate	ncases	Optimized	Shewhart	Improvement vs. Shewhart
1/14	200	0.798	0.798	0.00%
1/14	400	0.846	0.845	0.04%
1/14	1000	0.972	0.963	0.98%
1/14	2000	1.000	1.000	0.02%
1/28	200	0.556	0.555	0.08%
1/28	400	0.638	0.624	2.26%
1/28	1000	0.912	0.853	6.89%
1/28	2000	1.000	0.997	0.31%
1/56	200	0.346	0.339	1.92%
1/56	400	0.440	0.404	9.04%
1/56	1000	0.821	0.675	21.51%
1/56	2000	0.999	0.980	1.98%
1/112	200	0.203	0.191	6.05%
1/112	400	0.288	0.240	20.00%
1/112	1000	0.712	0.484	47.04%
1/112	2000	0.998	0.932	7.15%

Table 4-6: Optimal Detection Weightings for Late-peak Lognormal
The improvement found when optimizing detection of a lognormal outbreak which peaks on the fourteenth day. The number of total additional cases due to the outbreak is indicated in the ncases column.

Even when optimizing for a specific false alert rate, the optimized weighting often shows improvements over a range of false alert rates. This is shown in Figure 4-12. For this outbreak of 1000 cases, optimizing for a false alert rate of 1 per 28 days results in improvement over nearly all false alert rates. This can be seen in the table,

noting that the optimized weighting for the 1000-case outbreak is fairly constant (the improvement for a 1/28 FA rate is less than 0.00 percent, thus is not highlighted).

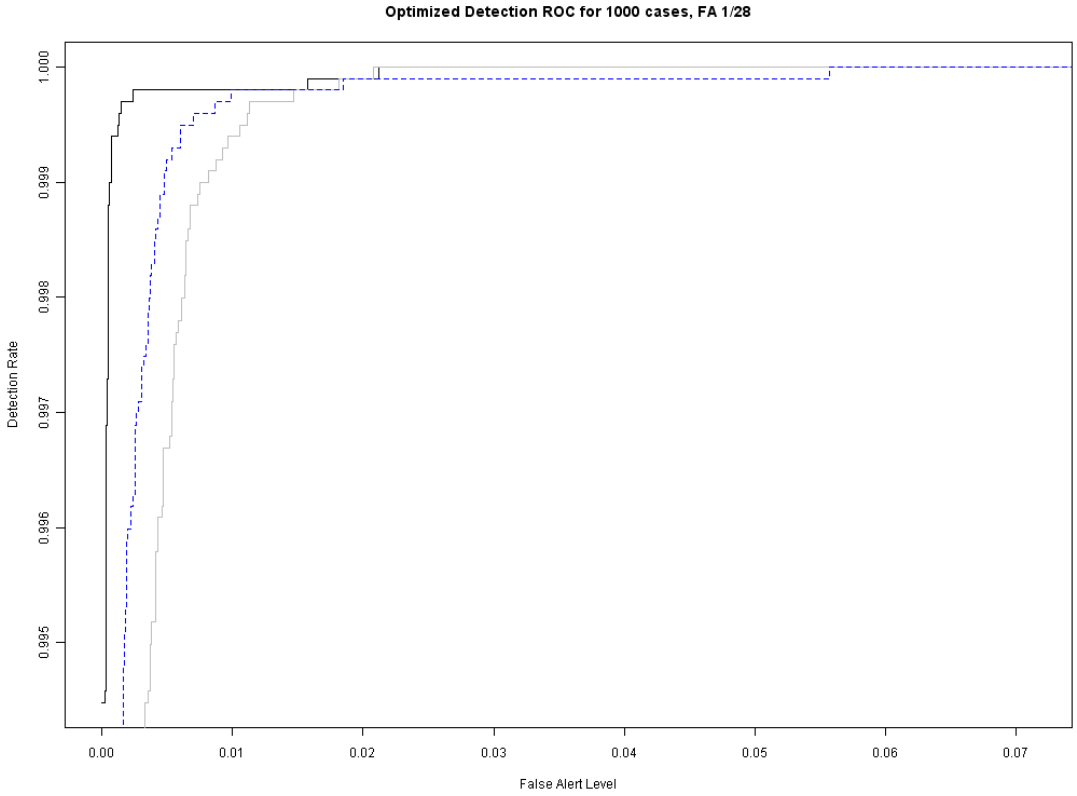

Figure 4-12: ROC for Optimized Detection on Lognormal Outbreak
The corresponding ROC curve (showing only the portion up to false alert level 1/14) for optimizing detection of a 1000-case lognormal outbreak which peaks on day 4. The optimized method is in black, the Shewhart in grey, and the CuSum in dashed blue.

Now that we have optimized for detection, we can apply the technique to optimize for timeliness. We can use the same direct solution methodology to calculate the expected delay for a given set of weights. Using this, we can optimize to find the weighting which provides the earliest mean detection day. Table 4-7 displays the results for some sets of false alert rate and outbreak size. We can see that the percentage improvement is generally lower than when comparing overall detection; when the Shewhart detects an outbreak, it has a good chance of detecting early. More

importantly, we can see that when optimizing for timeliness, the tendency towards the CuScore weightings is much weaker. As compared to the detection optimization, the weightings are closer to the Shewhart, reflecting the CuScore's focus on last-day detection. Similar to the detection scenario, the tendency towards multi-day weightings comes in regions of low false alert levels, for weaker outbreak signals (the 10-case outbreak is too small to reflect the lognormal shape very strongly). In these cases, detecting on the last day can make a significant improvement to timeliness. This supports the conclusion that having multiple days to detect results in non-CuScore methods being optimal for detection and for timely detection.

FA Rate	ncases	weight1	weight2	weight3	Improvement vs.Shewhart
1/14	10	0.00	0.00	1.00	
1/14	100	0.00	0.00	1.00	
1/14	1000	0.00	0.00	1.00	
1/14	10000	0.00	0.00	1.00	
1/28	10	0.00	0.00	1.00	
1/28	100	0.00	0.08	0.92	0.03%
1/28	1000	0.00	0.04	0.96	0.03%
1/28	10000	0.00	0.00	1.00	
1/56	10	0.00	0.00	1.00	
1/56	100	0.00	0.16	0.84	0.09%
1/56	1000	0.00	0.08	0.91	0.16%
1/56	10000	0.00	0.00	1.00	
1/112	10	0.00	0.00	1.00	
1/112	100	0.04	0.19	0.77	0.11%
1/112	1000	0.00	0.12	0.88	0.42%
1/112	10000	0.00	0.00	1.00	

Table 4-7: Optimal Timeliness Weightings for Lognormal
This table shows weights for optimization of the earliest mean day of detection. Rows with an improvement over Shewhart are highlighted.

4.4.6. Empirical Results

To confirm that these results apply to real health data sets, we used the same technique to determine the optimal weighting, but applied it to residuals from a Holt-Winters forecast on respiratory health series data (described in Section 1.3.1). The

results are contrasted with a standard Shewhart applied to the Holt-Winters residuals. Because the residual standard deviation is higher than in the previous example (approximately 65, compared to the 40 used for simulation), the optimized weights are slightly different. However, we can see that the pattern of improvement is virtually identical to that found in simulated data. For larger outbreak sizes, particularly for lower false alert levels, the optimized-weighting detector results in a marked improvement over Shewhart in overall detection, and a slight improvement in timeliness. This can be seen in Table 4-8 and Table 4-9. Although it is expected, it should also be noted that the optimization method, even though it is optimizing over an idealized case, never results in a weighting which performs worse than the Shewhart. Tests on later-peaking outbreaks show significantly improved performance, in line with performance predicted by the analysis in Section 4.4.5.

FA Rate	ncases	weight1	weight2	weight3	% improvement
1/14	20	0.000	0.000	1.000	
1/14	50	0.000	0.000	1.000	
1/14	100	0.000	0.000	1.000	
1/14	200	0.000	0.000	1.000	
1/14	400	0.000	0.104	0.896	0.18%
1/28	20	0.000	0.000	1.000	
1/28	50	0.000	0.000	1.000	
1/28	100	0.000	0.000	1.000	
1/28	200	0.000	0.106	0.894	0.08%
1/28	400	0.002	0.207	0.791	0.72%
1/56	20	0.000	0.000	1.000	
1/56	50	0.000	0.000	1.000	
1/56	100	0.000	0.096	0.904	0.02%
1/56	200	0.000	0.182	0.818	0.16%
1/56	400	0.034	0.248	0.718	1.04%
1/112	20	0.000	0.000	1.000	
1/112	50	0.000	0.064	0.936	0.00%
1/112	100	0.011	0.156	0.833	0.03%
1/112	200	0.034	0.227	0.740	0.17%
1/112	400	0.085	0.267	0.647	1.10%

Table 4-8: Optimal Timeliness Weightings on Authentic Data

This table shows, on authentic health respiratory data, the optimized weightings found for detection timeliness, using the method described in Section 4.4.4. Combinations of false alert rate and outbreak size which resulted in an improvement over the Shewhart are highlighted.

FA Rate	ncases	weight1	weight2	weight3	weight4	% improvement
1/14	20	0.000	0.000	0.000	1.000	
1/14	50	0.000	0.000	0.000	1.000	
1/14	100	0.000	0.000	0.000	1.000	
1/14	200	0.000	0.000	0.072	0.928	0.19%
1/14	400	0.000	0.000	0.214	0.786	2.09%
1/28	20	0.000	0.000	0.000	1.000	
1/28	50	0.000	0.000	0.000	1.000	
1/28	100	0.000	0.000	0.086	0.914	0.13%
1/28	200	0.000	0.012	0.174	0.813	2.12%
1/28	400	0.000	0.056	0.268	0.677	7.10%
1/56	20	0.000	0.000	0.000	1.000	
1/56	50	0.000	0.000	0.038	0.962	0.05%
1/56	100	0.000	0.000	0.153	0.847	1.28%
1/56	200	0.000	0.042	0.231	0.727	5.57%
1/56	400	0.000	0.063	0.288	0.649	14.69%
1/112	20	0.000	0.000	0.031	0.969	0.01%
1/112	50	0.000	0.018	0.136	0.846	0.47%
1/112	100	0.032	0.008	0.201	0.760	2.79%
1/112	200	0.080	0.084	0.365	0.471	4.06%
1/112	400	0.004	0.133	0.318	0.545	24.24%

Table 4-9: Optimal Detection Weightings on Authentic Data

This table shows, on authentic health respiratory data, the optimized weightings found for overall detection, using the method described in Section 4.4.4. Combinations of false alert rate and outbreak size which resulted in an improvement over the Shewhart are highlighted.

We can also see the effect in the ROC curves for optimized Detection Rates. The optimization's improved performance holds not only for the specific false alert rate, but for a variety of different potential false alert rates. Indeed, if one looks across the optimized false alert rates with significant improvement (for example, with an outbreak effect size of 400) then one can see that the weightings are quite similar.

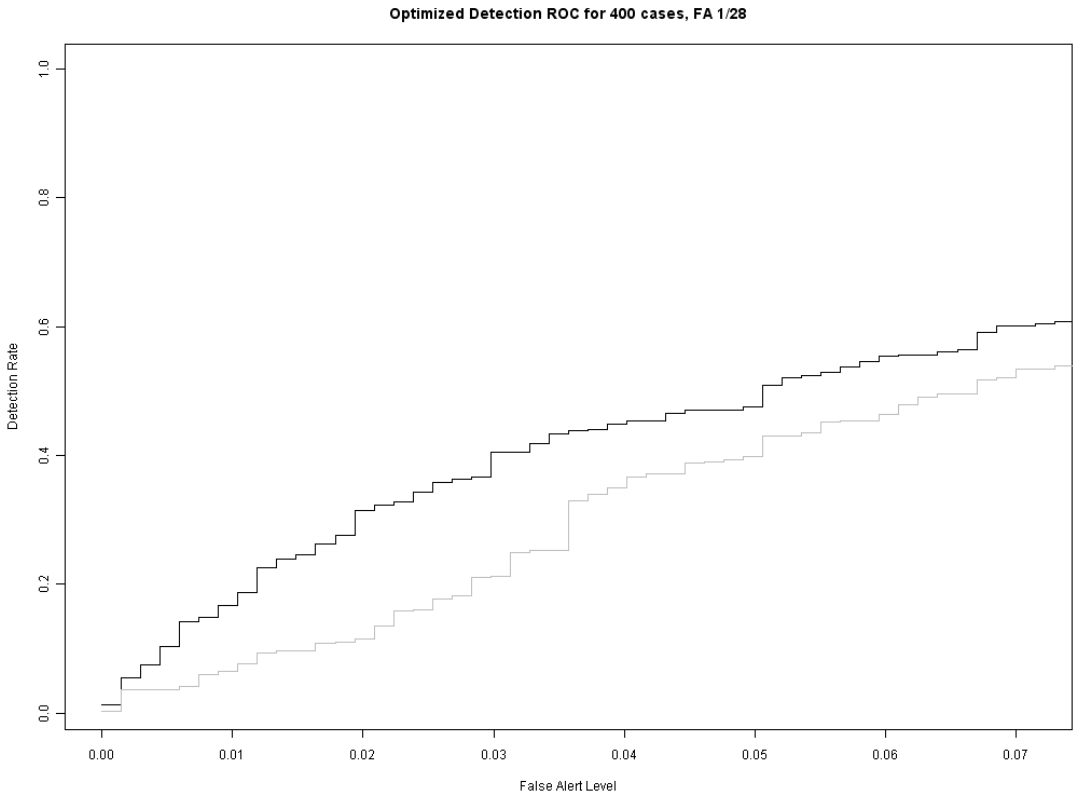

Figure 4-13: ROC for Optimized Detection on Multiple FA Levels
The corresponding ROC curve (showing only the portion up to false alert level 1/14) for optimizing detection of a 400-case lognormal outbreak. The optimized method is in black, the Shewhart in grey.

4.4.7. Conclusions and Future Work

The issues which have come up in the use of unmodified CuScore techniques show serious problems with using this technique for timely detection of multi-day anomalies. Because they do not consider the multiple opportunities for detection, they will not, on their own, provide improved detection. However, the optimization technique based on the CuScore weighting method is an effective method for detecting outbreaks with known shapes. Further, this should point the way for future work using a foundation of statistical theory to create a real improvement for practical outbreak detection.

In terms of future work, the optimization technique described in Section 4.4.4 could also be used to solve more general problems. For example, the technique could be used to maximize the probability of detection by a given day before the peak, in case of a response which needs to occur earlier in the progression of the disease to be effective. More generally, it could use an overall cost function for missed outbreaks, outbreaks detected on each day of the outbreak, and false alerts. This could impact the policy decisions of public health officials in deciding what diseases to focus on, and how many resources are required for various detection capabilities.

If the health series residuals significantly deviate from normality, one could modify the optimization method to use an alternative distribution (such as Poisson). If no parametric distribution is a good fit, then one could estimate probabilities from an empirical distribution, use Monte Carlo methods, or set up a Markov chain computation for estimating the probability of exceeding the control limit.

Finally, the analysis presented here assumes a fixed outbreak signal; as we showed in Chapter 2, having a stochastic outbreak signal can actually have a significant effect on performance. While the fixed outbreak is useful, it is ultimately an approximation. One easy and very useful extension to this work would be to consider a stochastic outbreak and make the corresponding changes to the daily detection analysis.

Chapter 5 : Improved Evaluation Methods

5.1. Introduction

Although the field of biosurveillance has grown in importance and emphasis in the last several years, the research community involved in designing and evaluating monitoring algorithms has not grown as expected. One reason for this lack of sufficient growth is the lack of publicly available data which researchers can use for developing and evaluating algorithms. Another reason is that the evaluation of different surveillance algorithms is done internally by each research group, thereby hindering open scientific evaluation of newly developed algorithms. One solution to this is to use simulated data, as described in Section 1.4.7. However, in order to be confident in the results form such simulation, one must be confident that the simulated data is similar to authentic data. For this reason, we present a way to apply statistical tests to evaluate simulated data.

In addition, when evaluating a detection algorithm or comparing two detection algorithms, an evaluator often has a variety of concerns. They are not simply concerned with the overall detection rate--they may be concerned with detection within the first 3 days *and* detection within the first 7. They are likely to be considering the benefits and costs from a variety of false alert rates as well. Because this information is not traditionally conveyed in a form which allows the health practitioner to consider multiple possible scenarios, we present a new visualization to display this information in an effective way.

5.2. Evaluating Simulation Effectiveness

A crucial component of using simulation to mimic authentic data is verifying that the simulated data retain the key characteristics of the original data. This is done by testing whether the simulated data come from the same distribution as the original authentic data. If they come from the same distribution, then the simulation method should be trustworthy and provide valid results; if not, then the differences between the original and simulated data can provide distorted and unrealistic results. To determine if this is the case, we present distribution tests specifically tailored for use in evaluating simulated biosurveillance data. These tests will also be published as part of (Lotze et al., 2010).

Of course, given a finite amount of original data, there exist an infinite number of distributions which could generate those data. The distribution tests presented here merely attempt to confirm that the simulation method is within that space of possible models, specifically those which have a reasonable chance of generating the data. We must use domain knowledge (such as our awareness of which characteristics are relevant) to further constrain the possible simulation models. Goodness-of-fit tests of the simulated data should be considered as relative measures of consistency; it is known that distributional tests become extremely sensitive with large amounts of data, and so may reject even the most useful simulations.

Finally, a mimic method will only be useful if it accurately captures the randomness of the underlying distribution. If a mimic is simply a duplicate of the original data, it

is clearly not a good additional test, nor does it avoid any privacy concerns. Similarly, a mimic which merely adds random noise to the original is not providing a new authentic set of possible data--it is simply providing the original data with extra variation.

5.2.1. Univariate χ^2 Testing

The first method for evaluating the closeness between the distribution of authentic and mimic data is a series of simple χ^2 tests. To test a simulated data set against its original data set, we take each univariate data series and split it by day of week. The values for a single day of week are then grouped into bins; an example of the binning process is given in Figure 5-1. The width of the bin varies by density, such that there are at least 10 observations in each bin. The original data are split and binned in the same fashion, and these two sets of counts (mimicked and original) are tested for distributional equality using a χ^2 test (with degrees of freedom equal to k-1, where k=the number of bins). An FDR (Benjamini & Hochberg, 1995) significance correction is used to account for multiple testing across multiple series. The χ^2 tests can also be repeated for each day-of-week separately with FDR correction, to inform us not only whether there are issues with our simulation, but also to point us towards the reasons for those issues.

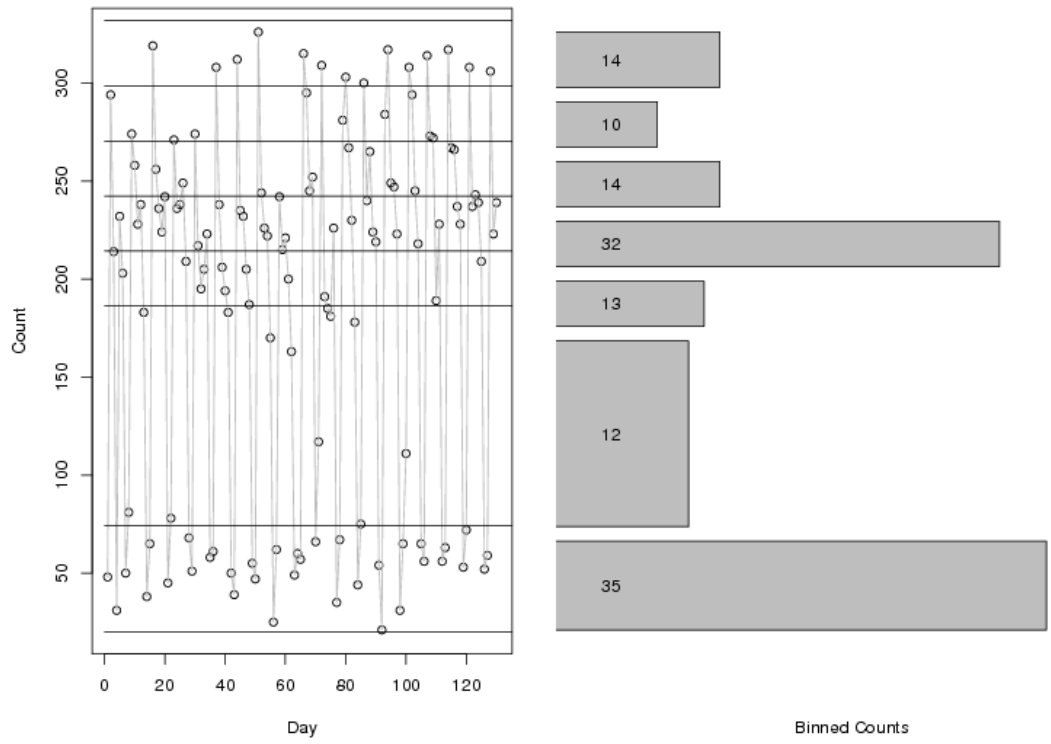

Figure 5-1: Binning of Simulated Time Series
A portion of a single time series being binned.

5.2.2. Multivariate Testing

The above χ^2 tests can only uncover univariate disparities between the original and mimicked data. To also consider the covariance between the series, we consider multivariate goodness-of-fit tests. While it is not obvious that such a test can be performed in a distribution-free manner, several methods have been developed to do so, notably (Bickel, 1969, Friedman & Rafsky, 1979, Schilling, 1986, Kim & Foutz, 1987, Henze, 1988, Hall & Tajvidi, 2002).

We use the nearest-neighbors test described in (Schilling, 1986), because of its asymptotic normality and computational tractability. Under this test, the nearest k

neighbors are computed for the combined sample. Each of the nearest neighbors is then used to determine an indicator variable, whether or not it shares the same class as the neighboring point. The statistic T, the proportion of k-nearest neighbors sharing the same class, is used to test equality of distributions. If both samples have the same size and come from the same distribution, T will approach 0.5 as the sample size increases. If the two samples differ in distribution, then T will tend to be larger than 0.5. With an appropriate correction, T has an approximate standard normal distribution. For an example, see Figure 5-2.

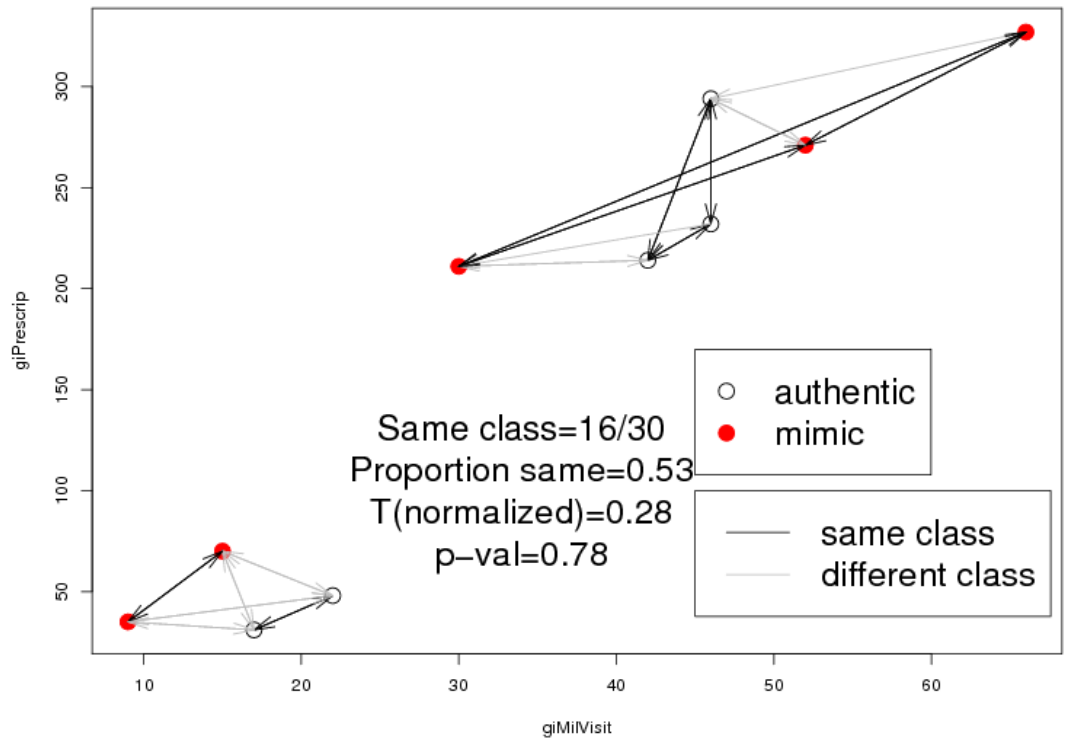

Figure 5-2: KNN Test
A simple example of the KNN test for multivariate distribution equality, using only two series and 5 time points from authentic and mimic series. Each point is labeled as authentic or mimic; the 3 nearest neighbors are computed, and an arrow is drawn connecting each point to its 3 neighbors. The line is black if the neighbors have the same label, grey if different. The number of neighbor links which are the same is summed, then normalized, and finally tested. Here, there is insufficient evidence to reject the null hypothesis, so we conclude that the authentic and mimic distributions may be the same.

5.2.3. Distribution Testing Example

We now consider the tests of distributional equivalence on the simulation method presented in (Shmueli et al., 2007). Simulated data was created to mimic the statistical properties of a city from the BioALIRT data set, providing 700 days of 6 time series. The multivariate nearest-neighbor test gives a Z-score of 3.62, with a p-

value of 0.000293. These p-values should be viewed cautiously, because due to the large sample size of n=1400 (700 for the authentic data and 700 for the simulated data), it will be very sensitive to any differences in distribution. Still, the value is quite low, leading us to consider the univariate χ^2 tests.

When individual day-of-week scores are considered for each series, we find significant deviations in four categories: giMilVisit on Sun (p-val=0.000915); giMilVisit on Sat (p-val=0.000225); giPrescrip on Sun (p-val=0.000045); and giCivVisit on Sun (p-val=0.000060).

Examining individual bin comparisons, we see that the mimics have less variance on weekends than the original, suggesting that a negative binomial with increased variance might improve the simulation method. Figure 5-3 shows differences in Sundays for GI Civilian visits.

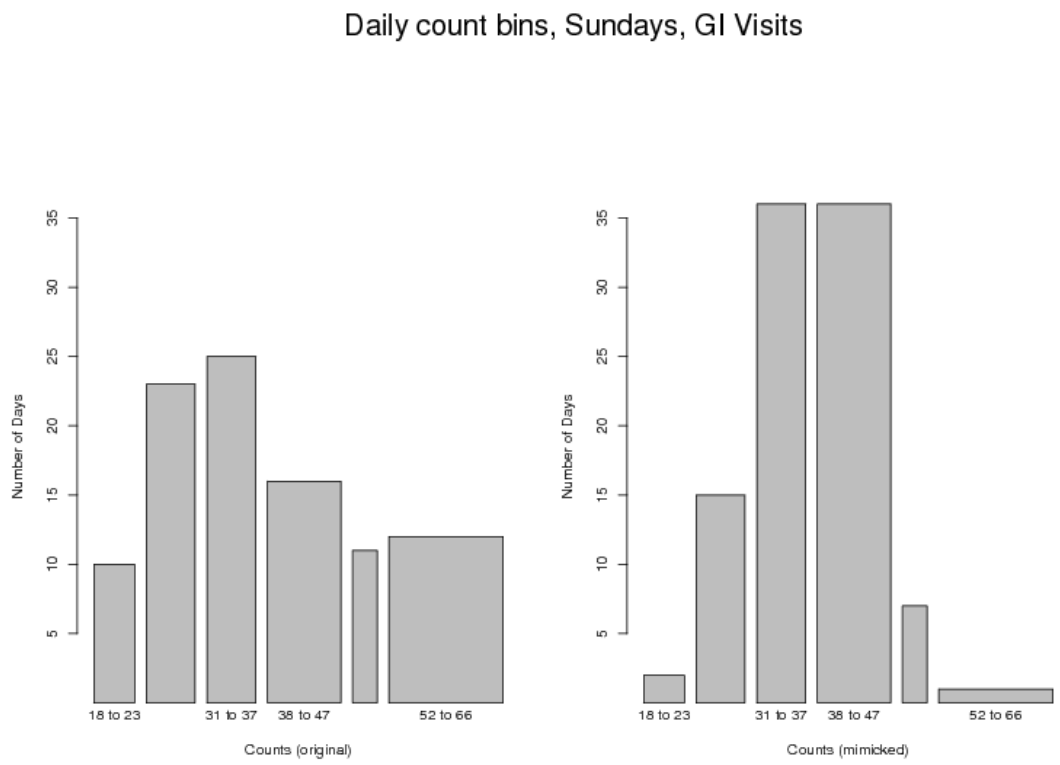

Figure 5-3: Chi-Squared Bin Test
An indication of a difference between authentic and mimicked data: the mimicked series tend to have lower variance than the authentic data.

This example shows how one might use these tests to find an issue with a simulation method, which could then be corrected to create an improved simulated data set. In addition, one could use these tests to compare multiple simulation methods against the same authentic data series, in order to rank them according to how well they capture the qualities of the authentic data.

5.3. Visualization

5.3.1. Problem Description

Aside from ensuring that results from different algorithms are numerically comparable and practically significant, it is important to convey that difference to a

researcher or to a practitioner who is evaluating different algorithms for use. In order to do this, one must be able to show the probability of detecting an outbreak within a certain number of days. This number of days will depend on the type of outbreak; one will be more concerned with early response to more virulent diseases, or ones for which earlier action is significantly more effective. It will also depend on the resources which it is being compared against, such as the speed with which lab results will show definitive signs of a disease outbreak. Being able to see this would provide researchers an effective way to judge algorithms against each other, and to provide practitioners a way of evaluating the practical usefulness of an algorithm.

5.3.2. Time-Lag Heatmaps

Rather than using ROC and AMOC curves to show detection probability and conditional timeliness for a given outbreak shape and size, we have developed a method to display the cumulative probability of detection on each day. We refer to this visualization as a time-lag heatmap. A time-lag heatmap displays, for each day, the probability that the algorithm will alert on that day or earlier during the outbreak. In this visualization, each row is interpreted as the series of days after the beginning of the outbreak (or, more accurately, after the beginning of the outbreak signal in the health series data). When the number of entries is small enough, the value of the actual probability is shown within each cell. An example is shown in Figure 5-4. In this example, the probability of detection by days 1 through 20 is shown. The probability in each cell is shown both by the darkness of the cell and by the numerical value.

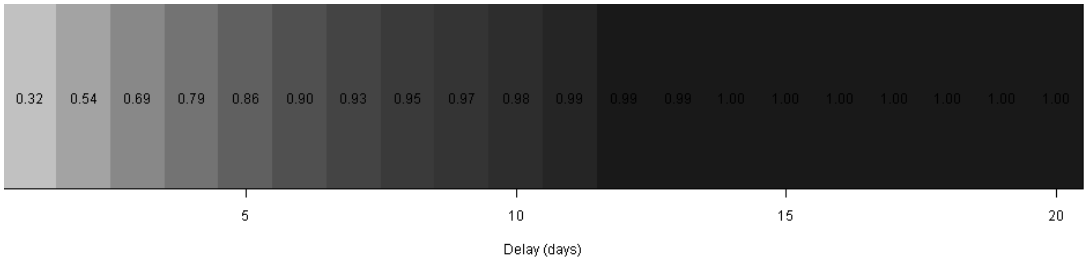

Figure 5-4: Cumulative Detection Probability Strip
This visualization shows the cumulative detection probabilities for a Shewhart detection algorithm on a 1-sigma step outbreak, with a false alert rate of 1 every 14 days. As should be expected, the detection probability goes to 1 as time progresses. The change in cell shading and numerical values show the cumulative probability of detection for each day.

In order to show the performance over a range of false alert levels, we generate a series of daily strips, one for each false alert level. By doing so, we can see how performance changes with different false alert rate requirements. An example is shown in Figure 5-5, which shows the probability of detection by day 1 through 20 for a given false alert rate as a horizontal strip. For example, the probability of detection by day 3 is 0.25 if we allow a 1/100 False Alert rate, but 0.35 if we allow a 1/56 False Alert rate. When color is available, the information can be even more effectively displayed, as in Figure 5-6.

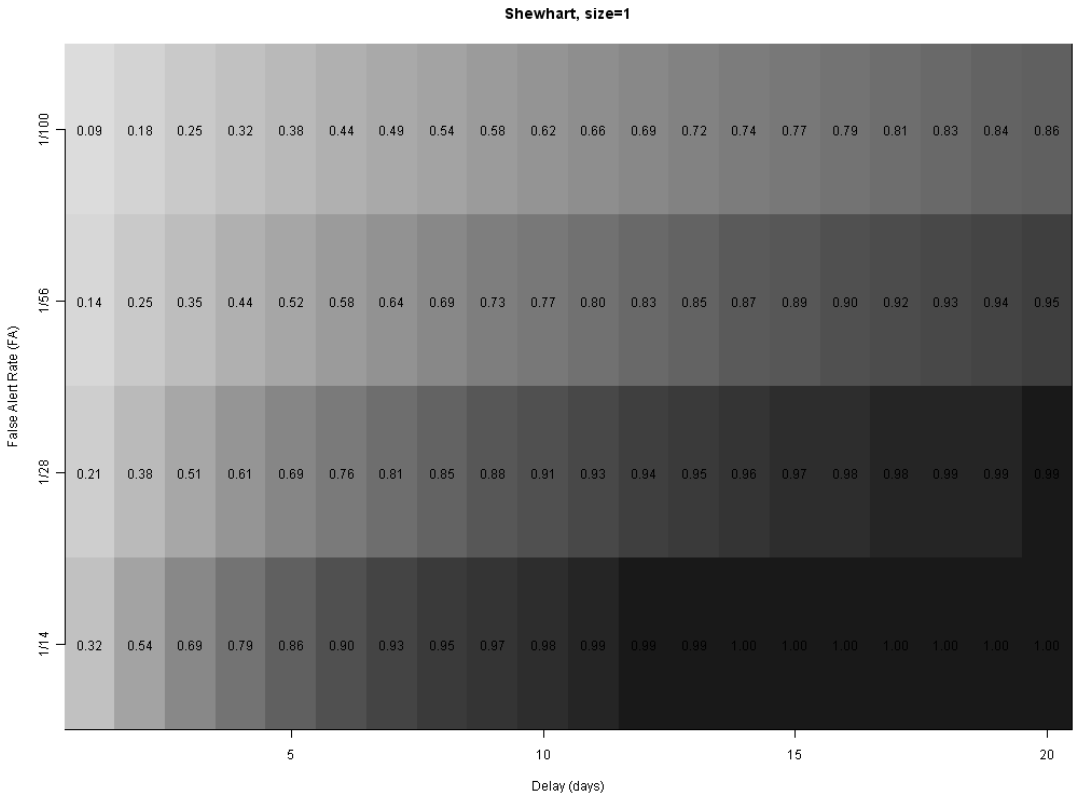

Figure 5-5: Time Lag Heatmap for Shewhart

This time-lag heatmap shows the probability of detection delay (for detection within the first 20 days) as the false alert rate decreases. The cumulative probability of detection for each day is on the x-axis, with different false alert levels on the y-axis, using a Shewhart chart on a 1-sigma step outbreak.

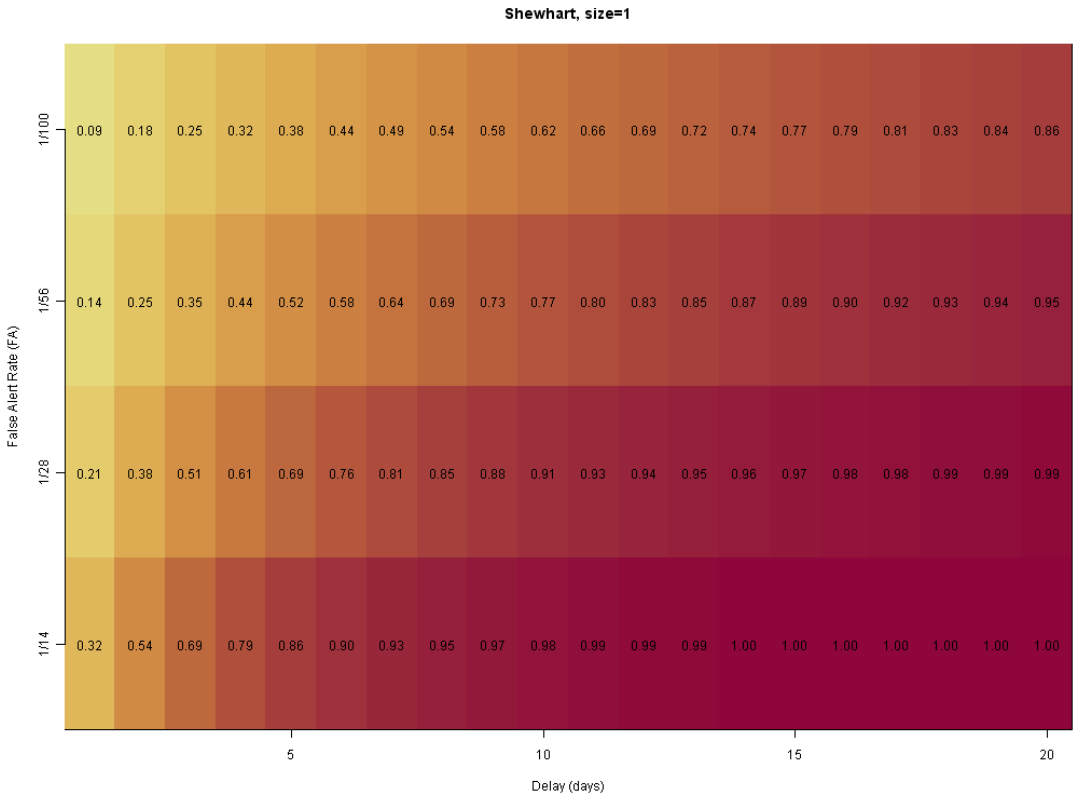

Figure 5-6: Time Lag Heatmap (color)
This time-lag heatmap shows the probability of detection delay (for detection within the first 20 days) as the false alert rate decreases. The cumulative probability of detection for each day is on the x-axis, with different false alert levels on the y-axis, using a Shewhart chart on a 1-sigma step outbreak.

We could also consider displaying this same information as a table of probabilities, without coloring the cells in the table as a heatmap. However. by coloring the individual cells for their cumulative detection probability, the visualization can be rapidly assessed and intuitively understood. While for smaller tables (ones with fewer days of interest or false alert levels), the shading is less beneficial, for tables comparing larger numbers of days or false alert levels, shading makes it possible to display and interpret what would otherwise be an unreadable table of dozens or hundreds of probabilities. Performance studies on visualization have shown effects on comprehension and speed from the method of visualization as well as its

organization (Henry & Fekete, 2006); while we have not performed a user study of the effect, we expect that this visualization method should improve the speed of comprehending the performance of different detection algorithms.

The time-lag heatmap visualization allows the display of both the detection performance and timeliness in one graph. For finite-time outbreaks, one can see not only the probability of detection (by reading the column for the maximum day of detection value), but also the probabilities of detection for any days prior. Thus, the time-lag heatmap includes the information from both the ROC and AMOC in a single graph.

5.3.2.1. Distribution over days

The time-lag heatmap method can also be used to show individual probabilities of detecting an outbreak on each day; thus, instead of showing the cumulative probability of having detected the outbreak, darkness represents the probability of detection on that specific day, highlighting days during which there is a higher probability of detection. This type of visualization can be used to show the detection probability distribution over days, which will often be more useful than simply reporting the mean or median day of detection.

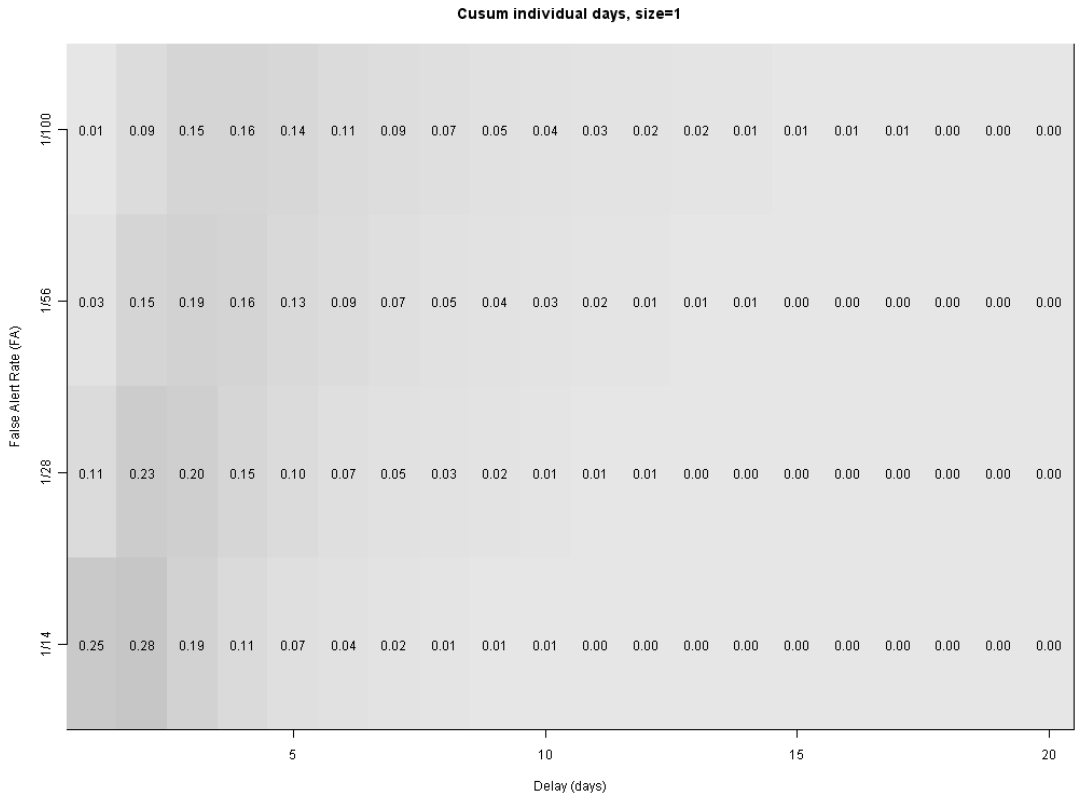

Figure 5-7: Individual Daily Detection Probability Heatmap
Individual daily detection probabilities for a CuSum chart, with differing false alert levels on the y-axis, for a 1-sigma step outbreak.

Using shading to denote the probability of detection on each day (not cumulative) we clearly see the shifting weight of the probable days of detection in Figure 5-7. By showing the distribution, we see both an increasing delay and an increased variance in the days of detection. This figure is also displayed more distinctly when color is available, as in Figure 5-8.

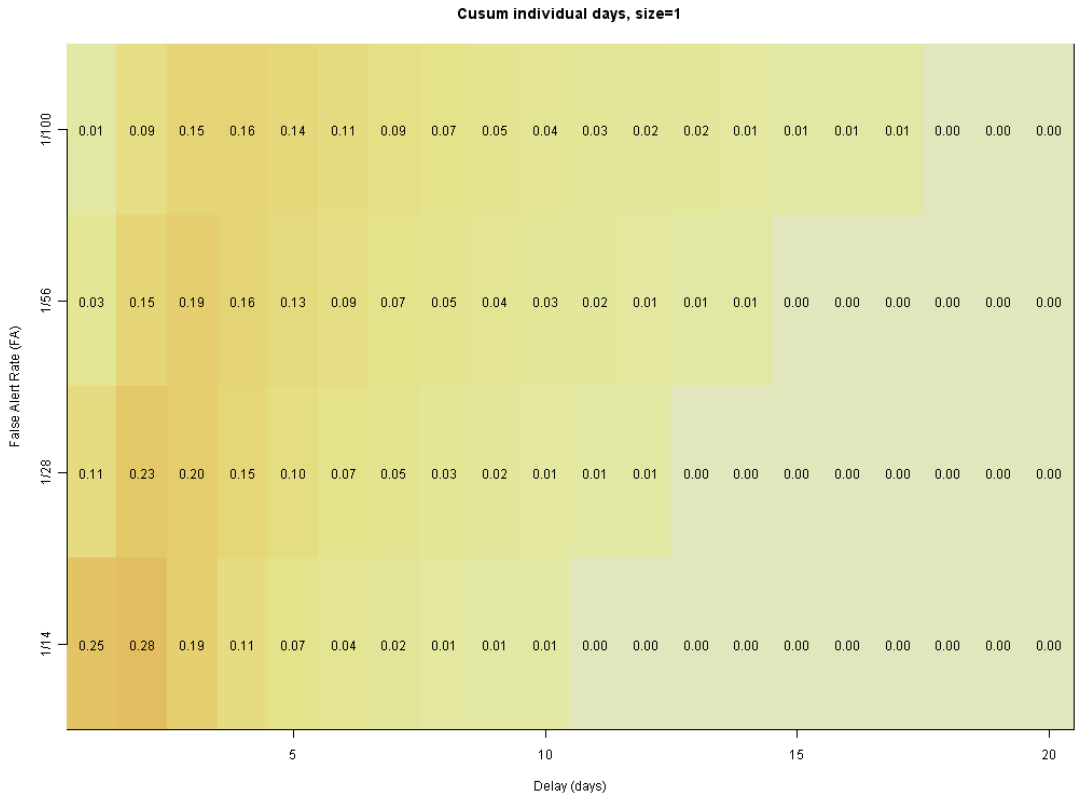

Figure 5-8: Individual Daily Detection Probability Heatmap (color)
Individual daily detection probabilities for a CuSum chart, with differing false alert levels on the y-axis, for a 1-sigma step outbreak.

5.3.3. Use in Evaluating Shewhart versus CuSum performance

The time-lag heatmap visualization depends on several components: the false alert rate, the nature of the underlying health data, and the outbreak type. Given these constraints, however, it can be used to illuminate some interesting and useful results. One such result involves the comparison of Shewhart versus CuSum control charts for detecting outbreaks.

We can compute the cumulative detection probability for each day for a Shewhart chart. Assuming a constant ATFS (False Alert Rate), we can compute the UCL (as described in Section 1.4.1) to be used as

$UCL = \sigma\Phi^{-1}(1 - \frac{1}{ATFS})$. Let p_i be the cumulative probability of detection on day i (i.e., the probability of detection on day i or earlier). For a Shewhart chart, the detection probability on day i is the right tail of the normal distribution, as it is simply the probability that the random variable is above the UCL. With an outbreak size for day i, given by η_i, the probability of detection is given by:

$$\mathrm{P}(\text{detection on day } i, \text{ given not detected before day } i) = 1 - \Phi(UCL/\sigma - \eta_i/\sigma)$$

Therefore P(detection on day 1)$= 1 - \Phi(UCL/\sigma - \eta_1/\sigma)$ and P(detection on day i) $= (1 - p_{i-1})(1 - \Phi(UCL/\sigma - \eta_i/\sigma))$. Thus, the cumulative probability of detection on day i is given by

$$p_i = \sum_{k=0}^{i} (1 - p_{k-1})(1 - \Phi(UCL/\sigma - \eta_k/\sigma)). \qquad \text{(Eq. 5-2)}$$

This quantity can be easily computed for any outbreak shape, using incremental computation for the p_i values. Recall Figure 5-5, which shows the time-lag heatmap for a Shewhart chart applied to a series with a step-increase outbreak shape.

For a CuSum chart, the probability of detection on each day can be calculated using Markov chain methods (Brook & Evans, 1972). However, using the same principles,

one can determine the cumulative or individual probability of detection. The results are shown in Figure 5-9.

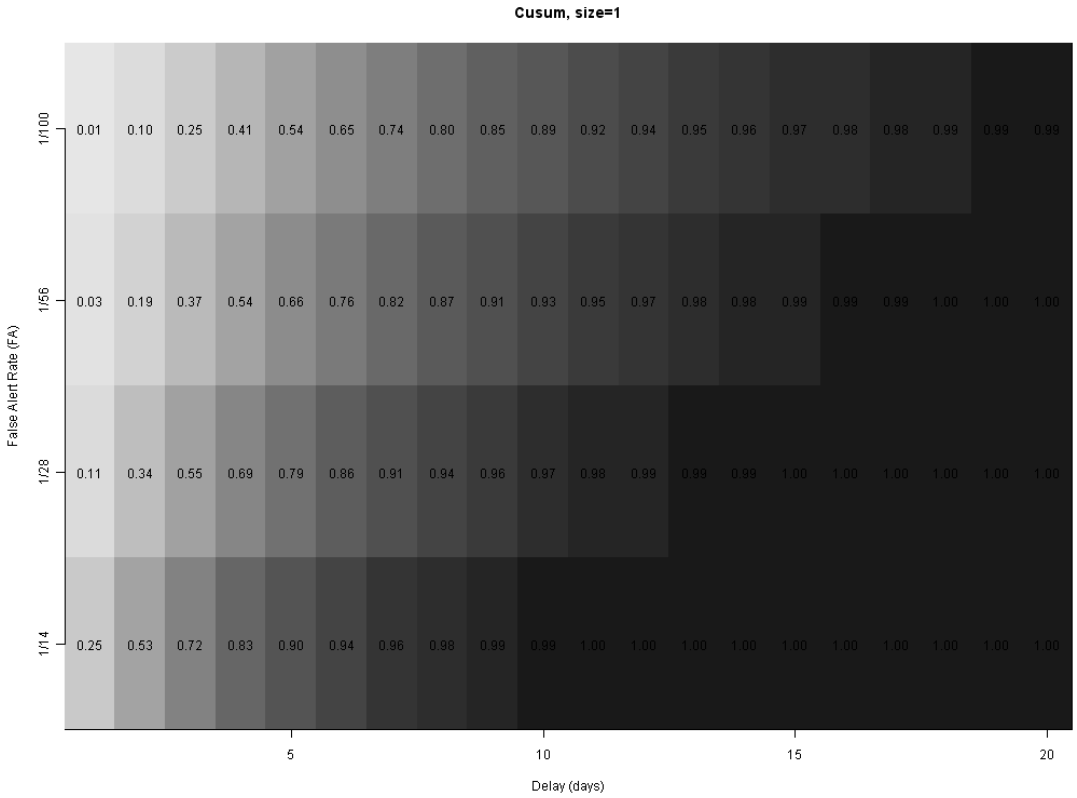

Figure 5-9: Time-Lag Heatmap for CuSum
This time-lag heatmap shows the cumulative probability of detection for each day on the x-axis, with different false alert levels on the y-axis, using a CuSum chart on a 1-sigma step outbreak.

To compare the performance of the Shewhart and CuSum charts, we can also generate a time-lag heatmap of their differences in cumulative detection probability, as seen in Figure 5-10. This is useful for examining the performance of a single algorithm under different false alert levels, or outbreak sizes, or for comparing two algorithms (e.g., Shewhart vs. CuSum). From this figure, we can see that while CuSum is better than Shewhart when the FA level is low (1/100, as used by (Kleinman & Abrams, 2006)), the differences are much smaller when the FA rate is

higher (e.g., 1/14). This resolves the apparent discrepancy between the theoretical analysis in Chapter 2 and reported results by (Kleinman & Abrams, 2006). When color is available, we can use a divergent HCL color scheme (Zeileis et al., 2009) to show different colors when the Shewhart or the CuSum is performing better. This version is shown in Figure 5-11.

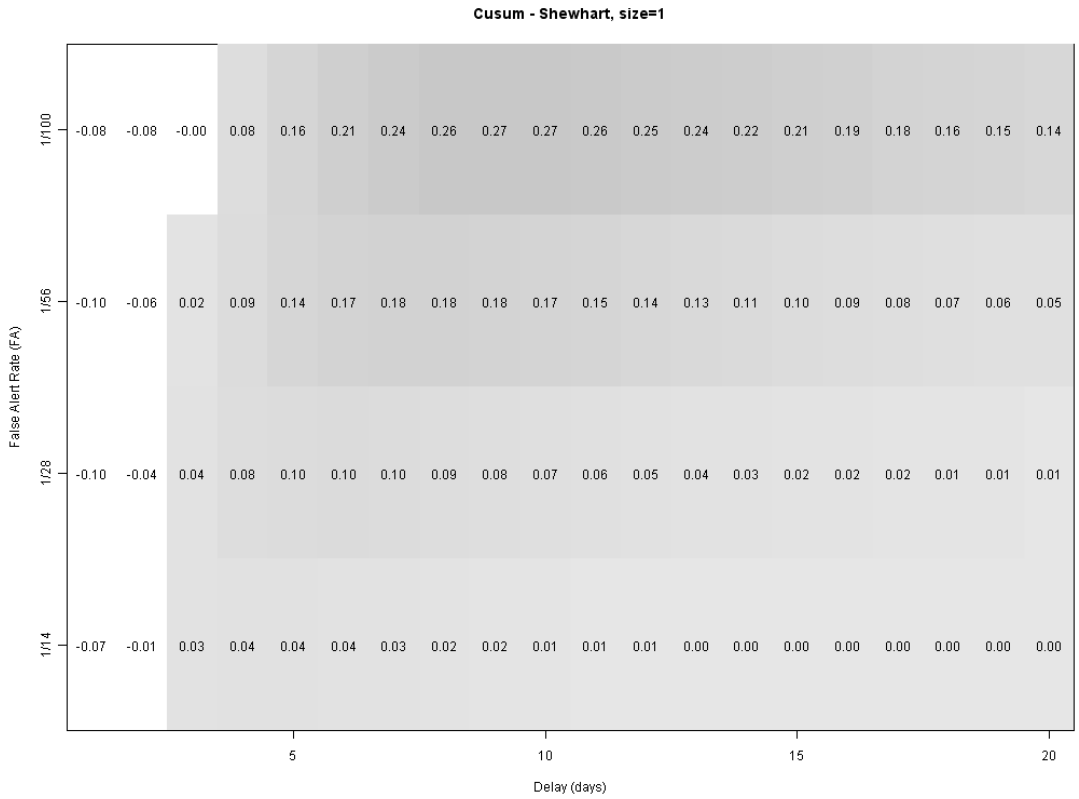

Figure 5-10: Time-Lag Heatmap for Difference Between Shewhart and CuSum
This shows the difference in cumulative probability of detection between CuSum and Shewhart detection methods. When the area is white (and the number negative), the Shewhart is performing better. The darker the area, the better improvement CuSum has over Shewhart, in terms of cumulative probability of detection.

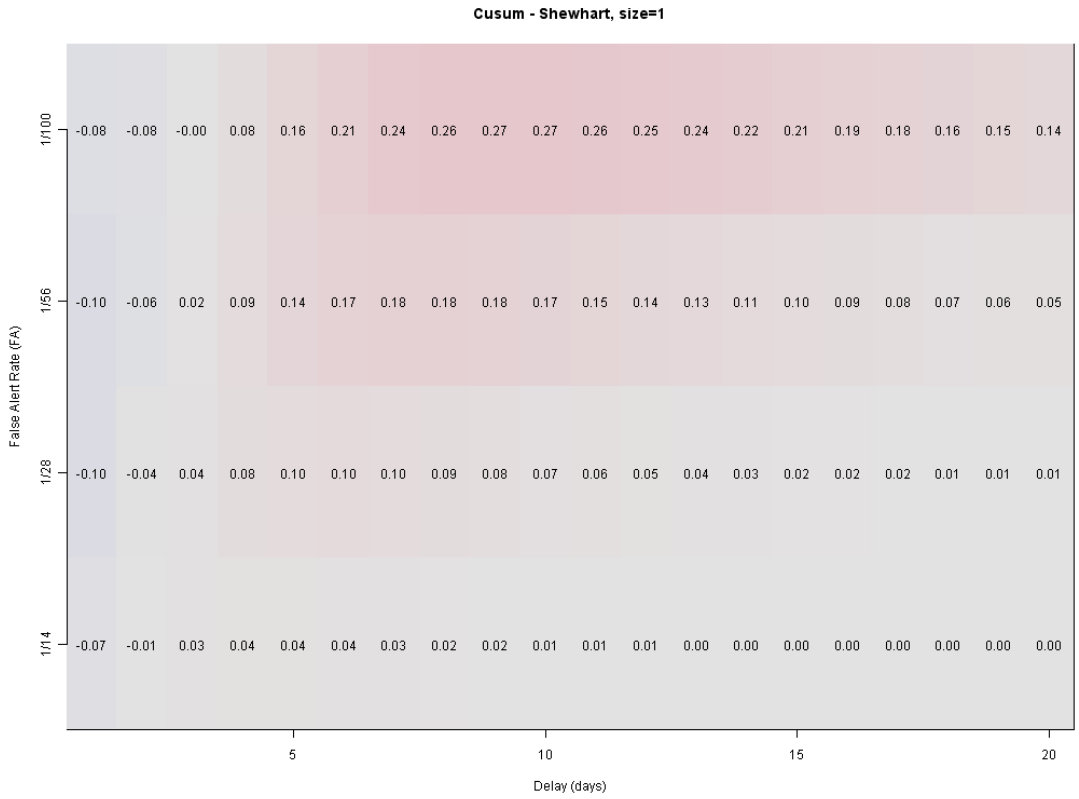

Figure 5-11: Time-Lag Heatmap for Difference Between Shewhart and CuSum (color)
This shows the difference in cumulative probability of detection between CuSum and Shewhart detection methods. When the area is more blue (and the number negative), the Shewhart is performing better. When the area is more red (and the number positive), the CuSum is performing better. The stronger the color, the stronger the difference; grey values indicate small differences.

5.4. Conclusions and Future Work

5.4.1. Simulation

An R package for mimicking multivariate time series and simulating outbreak functions is freely available at http://projectmimic.com, along with ten simulated data sets mimicked from an authentic biosurveillance data set. The R package is easily installed and contains extensive help for all functions, with example code. The data sets contain two years of data, with six health indicators from a single region. We encourage researchers to freely use the code or data sets provided. By creating

multiple simulated data sets which are "copies" of the same authentic data set, one can begin to investigate the sensitivity of an algorithm's performance to small variation, using randomization and Monte Carlo testing. The ability to test an algorithm on multiple versions of the same data structure helps avoid over-fitting and gives more accurate estimates of model performance.

By making the code and algorithms public and freely available, we hope to lower the barriers to entry and allow more researchers to become involved in biosurveillance. By providing a mechanism for generating mimics, we hope to encourage data holders to make mimics freely available. By providing a mechanism for testing mimics, we hope to evaluate methods for mimicking multivariate time series data and to improve such methods.

We believe that simulation can be an effective way of generating new, semi-authentic data sets for public research, free from privacy, confidentiality, and proprietary constraints. The tests presented here provide checks on the validity of the simulation, and allow us to consider further improvements in simulation of health data. By doing this, we hope to enable more researchers to consider the many challenges, and in particular statistical challenges, in biosurveillance (see (Shmueli & Burkom, 2009) for a survey of such challenges) and to provide an opportunity for rapid advancement of both research and practical solutions.

The evaluation tests considered here are unable to detect certain types of deviations between the authentic and mimicked data sets. For example, since the temporal factor is not considered, they will be unable to find differences in autocorrelation and other time-related deviations. For example, if all Saturday values were randomly reordered, the test results would be identical. Similarly, if the daily observations were reordered to have the same marginal distribution, but a different autocorrelation, this ordering would not cause a change in the test results. In addition, these tests will not find cases where the simulated data are too *close* to the original, such as when there is simple random variation around the original data points. As described above, however, this is an undesirable property of a mimic simulation. Tests for such scenarios should also be considered.

Ultimately, the best test of the mimicked data will be whether algorithms perform equally well on the mimicked data and on authentic data. If detection algorithms perform on authentic data as well as on mimicked data, we can be confident that our mimicked series are useful for testing and comparing algorithms. We can test this by simulating and injecting outbreak signals, then testing the performance of various algorithms on authentic versus simulated data.

5.4.2. Visualization

The time-lag heatmaps are an effective way to visualize the information contained in ROC and AMOC curves in a single graph. They provide a new visual representation which captures the most useful information for researchers and practitioners, and have direct interpretation in terms of detection probabilities. Time-lag heatmaps can

be modified to show many significant features of an algorithm's performance, or to compare two algorithm's performance, by highlighting the key feature of timely detection.

Additional modifications can be made to this basic idea in order to generate other types of useful graphs. One way of doing this would be to show a different factor on the Y-axis other than false alert rate or outbreak size. For example, it could vertically compare the performance of different algorithms applied to data with an outbreak of interest. The visualization method could also be extended by adding glyphs for additional information, such as the median or mean detection day. Finally, it would also be useful to perform a user study on the time-lag heatmap visualizations, to quantify the improvement in task completion when using them instead of other representations.

Chapter 6 : Conclusions and Discussion

6.1. Contributions of this Dissertation

In this dissertation, we have proposed a number of methods to improve algorithmic biosurveillance. First, we developed the theory for understanding the relationship between forecasting and detection. By doing so, we shed light on factors which affect an algorithm's detection performance; with that understanding, we can see where our existing algorithms are weak and improve upon them when possible. We have also started to investigate situations where improved forecasting will not result in improved detection. This theory has only started to be developed; but it has already explained several aspects of biosurveillance algorithm performance; future

developments should provide an even better understanding of the factors behind detection performance.

Second, we proposed methods to improve the forecasting of baseline health series. When multiple series are available, we can use cross-series covariates to provide additional information about the series of interest. The information can serve as a proxy for effects which are not directly measured, but which nonetheless affect the baseline behavior of the series. Similarly, we discussed the use of Temperature as one way of improving detection by using additional information which directly affects the behavior of the health series. By finding and incorporating additional sources of information such as temperature, we can also improve performance. Finally, we proposed an ensemble method for combining forecasters to improve forecasting performance. By combining multiple forecasters, we have the potential to create a forecaster which is better than any individual forecaster. By adding additional interaction effects, this could be further improved to allow an ensemble which uses different combinations of forecasters depending on how well they perform on different days or other aspects of the data. By improving forecasting, as we saw from the theoretical analysis, we can improve detection performance.

We have also proposed several methods to directly improve detection algorithms. The first of these is to combine multiple series into a single statistic to monitor, thereby providing improved performance by using the information from multiple sources. The second is a general-purpose method for normalizing the residuals

according to their estimated day-of-week variance. After doing this, the residuals are closer to having a common variance, and so show an outbreak in a more consistent way. This method can be easily applied to any method after forecasting and before applying detection, and has been shown to provide significant improvement over a wide range of outbreak sizes and false alert rates. Finally, we proposed and developed a new method, based on the CuScore, for finding optimal weighted detectors. This method allows one to find detectors which have the highest detection rate for a certain false alert rate and outbreak size. By using a normalizing approximation, we can find these optimal detectors quickly and easily. This method has been shown to have improved performance on real data, and can provide detectors which optimize overall detection, timely detection, or any cost function of detection on various days. These detection methods provide improved detection performance; in particular, the day-of-week standardization is an improvement which can be applied to a wide variety of detection algorithms, and the optimized detectors allow the ability to find the best detector, tuned to a particular outbreak signal.

Finally, we have proposed two ways to improve the evaluation of biosurveillance algorithms. First, we have developed two types of tests for evaluating simulated health data sets. By using these, researchers can find weaknesses in simulated data and improve the simulation methods to provide more useful test sets. These methods can also be helpful for improving the modeling of health sets, which should result in improved forecasting. Second, we proposed the use of Time-Lag Heatmaps for visualizing the daily detection probabilities of individual algorithms as well as for

comparing two algorithms. These visualizations provide an intuitive understanding of how well an algorithm performs, or where one algorithm outperforms another, as well as allowing for a quantitative comparison on individual days. This should allow researchers and practitioners to better understand the performance of different algorithms.

These improvements comprise a broad set of related improvements, working within the framework of improving biosurveillance by understanding the problem of anomalies in time series. By understanding the nature of this problem and comparing different methods, we can improve performance of the algorithms and so provide better tools to real public health practitioners.

6.2. Beyond Binary Detection

We define each day's problem as a binary detection question: is there an outbreak on this day? But while this formalization makes it possible for algorithms to solve the problem, there are two issues to consider. First, the binary setup provides a very coarse signal. Instead of simply indicating "outbreak" or "no outbreak", we should consider providing a measure of confidence along with the indications of outbreak. Because there is a range of possible strengths for outbreak indicators, this can help practitioners decide on the appropriate response. Some systems, such as ESSENCE, already rank potential outbreaks (Babin et al., 2008) or directly indicate the confidence in the alert as significant or mild (Burkom et al., 2008). Including confidence measures in the formal problem definition would make the systems more reliable for users.

A second weakness of the binary setup is that it is divorced from the question of action: what is the best response? In order to determine an appropriate course of action, one can think about two further questions: "What is the eventual size and shape of this outbreak going to be?" and "What type of disease and disease spread is being detected?" Currently, epidemiologists and other public health practitioners are responsible for determining the answers to these questions. If algorithmic approaches could provide additional insights, it could make them much more useful tools. It is possible that daily detection will not be sensitive enough to provide this kind of specific information. As information technology becomes increasingly integrated into health data providers and sentinel systems, information can be collected at an hourly level and eventually in real-time. Some organizations have already begun collecting health data in more frequent intervals (Wagner et al., 2006). But just as there were additional challenges when moving from weekly data, which are more consistent but also slower, so are there challenges in moving to more frequent data (Shmueli & Fienberg, 2006); algorithms will have to be adapted to high-frequency data and monitoring.

6.3. Confidence Intervals in Evaluation

In evaluating future biosurveillance algorithm results and comparing algorithm performance, we must consider confidence intervals and variance of the evaluation metrics. Point estimates and empirical averages alone cannot be relied upon to distinguish between methods' performance. If we do not provide an estimate of the variance of an algorithm's performance, then we cannot reliably say that it has

improved performance over another method. The issue can be mollified to some extent by simulation of additional data sets (thereby increasing the sample size for detection performance) or by a preponderance of evidence over multiple authentic data sets or outbreak simulations; but in order to claim a significant difference between performances, we should provide confidence intervals for that difference. While we are guilty of not including them in this dissertation, we recognize that in moving forward, this will be crucial to the future of biosurveillance research.

Because the statistical distribution of the evaluation metrics is not always known, research into these distributions could provide valuable understanding of when one method is significantly outperforming another. Even without a theoretical distribution, simulation to estimate the empirical distribution would be useful. In addition, for many of the methods described here, one can also provide confidence intervals; for example, treating the detection rate as the probability of a binomial distribution, and each simulated outbreak as a trial, one could provide binomial confidence intervals for the true detection rate or for the difference between two detection rates (including a multiple testing correction). Similarly, the methods described in Chapter 2 could also provide confidence intervals for the detection rate of an algorithm simply by recognizing that the detection will come from a binomial distribution. Finally, research into the effect of using the empirical false alert rate to set the upper control limit (rather than setting it in advance based on theoretical assumptions) would be quite useful for providing more accurate confidence intervals.

6.4. The Larger Context

The methods proposed here are mainly described in terms of early detection and automated alerts. The theory described can provide a better understanding of the factors related to performance, and the improved methods presented can provide better detection performance. However, it is important to recognize that these automated algorithms are a single tool in the toolbox. An automated alerting algorithm will not be the only indicator of a disease outbreak; but especially in combination with epidemiologist investigation, it can provide valuable insights into the current health situation and also give crucial corroborating evidence.

Early detection is a mechanism which can provide notification of a possible outbreak before it would otherwise have been noticed, so that an investigation can begin. But in practice, algorithm-assisted biosurveillance is both more and less than this. Automated algorithms are not reliable enough (partly due to the issues described at the end of Section 2.8) to be the sole determinant of a response: they provide too many false alerts and not enough true detections to justify a school closing or even a warning to all hospitals without further investigation. In addition, they are generally coarse tools which can detect several different indicators of outbreaks without identifying the specific disease or subpopulation which is affected. However, by searching through different possibilities and attempting to find areas of statistical significance, they can be a significant aid to professionals who want to find potential outbreaks, but cannot spend the hours needed to look at every possibility. They can provide good indications for broader and deeper investigation, investigation which

can result in a more specific and useful understanding of the cause of the outbreak and an effective response. Second, when there is already clinical suspicion, these tools can be used to provide quantitative validation of that suspicion, providing evidence that something is significantly different. By doing so, they give the practitioner a more convincing case and help them make a better decision on the correct response. Hence, detection algorithms serve as a decision support system rather than an independent alerting mechanism.

In real situations, algorithm detection systems have provided important indicators for further investigation as well as quantitative evidence of significantly increased cases due to outbreaks (CDC, 2007); but understanding and responding to the situation still requires trained professionals and expert analysis. These detection algorithms are valuable, and improving their performance is important, but neither the algorithms nor clinical knowledge is as effective alone as when the two reinforce and support each other. We must remember the larger context to improve public health response for real outbreaks.

Appendix A: Mathematical Notation

- e_t is the forecast error on day t, considering only the baseline (non-outbreak) health series: $e_t = u_t - f_t$
- f_t is the forecasted value for day t
- $\ln(x)$ is the natural logarithm of x: $\ln(x) = \log_e(x)$
- $[[o_i]]$ describes a vector $[o_1 o_2 ... o_n]$
- o_t is the outbreak signal for day t; it is 0 if there is no outbreak occurring on day t
- $\Phi(k)$ is the Gaussian cdf with mean $\mu = 0$, standard deviation $\sigma = 1$: $\Phi(k) = \int_{x=-\infty}^{k} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx$
- r_t is the residual value on day t: $r_t = y_t - o_t$
- σ is the standard deviation of a random variable
- u_t is the underlying baseline health data series count for day t
- y_t is the observed health data series count for day t: $y_t = u_t + o_t$
- Z is a standard Gaussian random variable with mean $\mu = 0$ and standard deviation $\sigma = 1$, Z~N(0,1)
- z_α indicates the one-sided upper Gaussian alpha quantile: $1 - \Phi(z_\alpha) = \alpha$ (equivalently, $P(Z > z_\alpha) = \alpha$)

Glossary

ACF: AutoCorrelation Function; see Autocorrelation.

AMOC: Activity Monitoring Operating Characteristic; an AMOC curve plots 1/ATFS (false alert rate) on the X-axis and ATFOS (average delay before detection) on the Y-axis. It displays a detection algorithm's timeliness over a range of false alert levels. See Section 5.1.

ARL: Average Run Length; this is a general term for a detection process' average time until it generates an alert. When the system is in control (or there is no outbreak), this is the ATFS. When the system is out of control (or there is an outbreak), this is the ATFOS. See Section 2.2.3.

ATFOS: Average Time to First Outbreak Signal; this is the average time after an outbreak begins until the detection algorithm provides an alert. It may also be called Delay, Timeliness, or Average Delay. See Section 1.1.4.

ATFS: Average Time to False Signal; when there is no outbreak, this is the average time until a detection algorithm generates an alert. 1/ATFS will often be referred to as the False Alert rate. See Section 1.1.4.

AUC: Area Under the Curve; this refers to the area under a ROC curve, and measures a detection algorithm's performance over a range of false alert levels. See Section 5.1.

Autocorrelation: A time series is autocorrelated if successive values (i.e., y_t and y_{t+1}) are correlated. This generally indicates that there is some common factor influencing nearby values, or that effects on the series have lasting impact. See Section 2.5.2.

Bernoulli: A Bernoulli trial is a trial with two outcomes, usually defined as success (1) or failure (0). It is essentially a weighted coin flip. See Section 2.2.3.

BioALIRT: Bio-Event Advanced Leading Indicator Recognition Technology; a project sponsored by DARPA to provide data and evaluate biosurveillance algorithms' ability to detect outbreaks in that data. See Section 1.3.1.

BioSense: A CDC biosurveillance program. See Section 1.2.2.

Chebyshev: Chebyshev's Inequality is a bound on the number of values in a sample or distribution which are far from the mean. It states that if the mean is μ and standard deviation is σ, then for any number k, at least $1 - \frac{1}{k^2}$ of the values are within $\mu \pm k\sigma$. See Section 2.3.

Covariance: The covariance of two random variables measures their linear relationship. $Cov(X, Y) = E((X - E(X)(Y - E(Y)))$. Higher covariance indicates that the two are more related: when one is high, so is the other. A covariance of 0 indicates no linear correlation. A negative covariance indicates variables which move in opposite directions (when one is high, the other is low). Because covariance is strongly affected by the variance of the individual variables, correlation is often used instead.

CuScore: A CuScore is a score designed to have maximum correlation with a particular signal. Monitoring a CuScore is a detection method used for detecting occurrences of a specific signal type. See Section 4.4.

CuSum: A common control chart method, which measures Cumulative Sums of deviations from an expected mean. See Section 1.4.1.

Delay: The amount of time after an outbreak begins until it is detected. See ATFOS.

Detection Probability: The probability a detection method has of detecting an outbreak. See Section 1.1.4.

EARS: Early Aberration Reporting System; a CDC biosurveillance project now included in BioSense. It defines several algorithms which are commonly used in practice or for comparison with new algorithms. See Section 1.2.2.

ED: Emergency Department. The number of people, each day, indicating a specific type of chief complaint (such as a respiratory problem) is a common source of biosurveillance data.

Efficient statistic: An efficient statistic is one which has minimum variance over all comparable statistics measuring the same underlying value on the same set of data. See Section 4.4.1.

ER: Emergency Room. See ED.

ESSENCE: Electronic Surveillance System for the Early Notification of Community-Based Epidemics; a biosurveillance program run by the Department of Defense and Johns Hopkins university Applied Physics Laboratory. See Section 1.2.3.

EWMA: Exponentially Weighted Moving Average; a common control chart method for monitoring a series. See Section 1.4.1.

FA: False Alert Rate; see ATFS.

Gaussian: The Gaussian distribution (often called the Normal distribution) is the familiar bell curve distribution. It is often used to approximate a random

variable's distribution, due to mathematical tractability and theoretical justification (many distributions will tend towards a Gaussian as larger amounts of data are observed).

Geometric: A geometric distribution describes the number of Bernoulli trials needed before the first success. See Section 2.2.3.

GI: Gastrointestinal. Relating to the stomach and/or small and large intestines. A category of chief complaint in ED data.

Heatmap: A visualization method in which the values are displayed as colors rather than numbers. See Section 5.3.2.

Holt-Winters: An adaptive forecasting method which uses a level, linear trend, and seasonal component. Sometimes referred to as HW or Holt-Winters Exponential Smoothing. See Section 3.2.3.

HW: See Holt-Winters.

ISDS: International Society for Disease Surveillance; a society which aims to advance the field of disease surveillance. It provides a forum for researchers and practitioners to work together, publishes a journal (Advances in Disease Surveillance), and hosts an annual conference.

MSE: Mean Squared Error; the average squared error.

$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{X}_i - X_i)^2$. See Section 2.1.2.

Normal: See Gaussian.

OTC: Over-the-Counter; the total sales, per day, of over-the-counter medication such as pain relievers or cough syrup is a common source of biosurveillance data.

Outbreak Signal: An outbreak signal is the expected number of additional cases due to a disease outbreak; it may also be generated by the expected delay from infection to display of symptoms. An outbreak signal is frequently added to a baseline data set to test whether an algorithm can detect it.

Poisson: The Poisson distribution is a common distribution of count data. It arises when one is measuring total number of events within a period of time, when there is an underlying common average probability of an event occurring, and events occur independently.

Regression: A method of relating an outcome variable to predictor values. It is commonly used in biosurveillance to forecast the value of a health series. While regression is commonly used as shorthand for linear least-squares regression, there are actually a variety of methods which are also called regression. See Section 3.2.1.

Resp: Respiratory. Relating to the lungs and/or airway. A category of chief complaint in ED data.

RMSE: Root Mean Squared Error; the square root of the average squared error. $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{X}_i - X_i)^2}$. See Section 2.2.2.

ROC: Receiver Operating Characteristic; a ROC curve plots 1/ATFS (false alert rate) on the X-axis and Detection Rate (true alert rate) on the Y-axis. It displays a detection algorithm's detection rate over a range of false alert levels. See Section 5.1.

RODS: Real-Time Outbreak and Disease Surveillance; a biosurveillance program created by the University of Pittsburgh. See Section 1.2.1.

Shewhart: A common control chart method which monitors the series directly. See Section 1.4.1.

SPC: Statistical Process Control, a field interested in monitoring processes for defects by using control charts.

TA: True Alert Rate; see Detection Probability.

Timeliness: See ATFOS.

Time series: A time series is a sequence of measurements or observations over time: $y_1, y_2, \ldots$. In biosurveillance, there is usually one value each day. An example time series might be the total number of cough syrup remedies sold, each day, in a particular geographic region. See Section 1.1.1.

UCL: Upper Control Limit; for a detection process, this is a value used as the upper bound for normal behavior. Any value above this limit is considered to be an anomaly (or out of control) and generates an alert.

Bibliography

[Adams *et al.*, 2006] Adams, B. M., Saithanu, K., & Hardin, J. M. 2006 (August). *A Neural Network Approach to Control Charts with Applications to Health Surveillance*. Invited talk at the 2006 Joint Statistics Meeting. Seattle, Washington.

[Archibald *et al.*, 1997] Archibald, LK, Manning, ML, Bell, LM, Banerjee, S, & Jarvis, WR. 1997. Patient density, nurse-to-patient ratio and nosocomial infection risk in a pediatric cardiac intensive care unit. *Pediatric Infectious Disease Journal*, **16(11)**, 1045–1048.

[Atienza *et al.*, 1997] Atienza, O.O., Ang, B.W., & Tang, L.C. 1997. Statistical process control and forecasting. *International Journal of Quality Science*, **2**, 37–51.

[Babin *et al.*, 2008] Babin, Steven M., Burkom, Howard S., Mnatsakanyan, Zaruhi R., Ramac-Thomas, Liane C., Thompson, Michael W., Wojcik, Richard A., Lewis, Sheri Happel, & Yund, Cynthia. 2008. Drinking Water Security and Public Health Disease Outbreak Surveillance. *Johns Hopkins APL Tech. Digest*, **27(4)**, 403–411.

[Bean & Martin, 2001] Bean, N H, & Martin, S. M. 2001. Implementing a network for electronic surveillance reporting from public health reference laboratories: An international perspective. *Emerging Infectious Diseases: Perspectives*, **7**, 773–779.

[Benjamini & Hochberg, 1995] Benjamini, Y., & Hochberg, Y. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B*, **57**, 289–300.

[Benneyan, 1998a] Benneyan, J. C. 1998a. Statistical quality control methods in infection control and hospital epidemiology, Part I: Introduction and basic theory. *Infection Control and Hospital Epidemiology*, **19(3)**, 194–214.

[Benneyan, 1998b] Benneyan, J. C. 1998b. Statistical quality control methods in infection control and hospital epidemiology, Part II: Chart use, statistical properties and research issues. *Infection Control and Hospital Epidemiology*, **19(4)**, 265–283.

[Bickel, 1969] Bickel, P.J. 1969. A Distribution Free Version of the Smirnov Two Sample Test in the p-Variate Case. *The Annals of Mathematical Statistics*, **40(1)**, 1–23.

[Bilenko *et al.*, 2003] Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., & Fienberg, S.E. 2003. Adaptive Name Matching in Information Integration. *IEEE Intelligent Systems*, **18**(5), 16–23.

[Blegen MA, 1998] Blegen MA, Vaughn T. 1998. A multisite study of nurse staffing and patient occurrences. *Nursing Economics*, **16(4)**, 196–203.

[Box & Luceno, 1997] Box, G., & Luceno, A. 1997. *Statistical Control: By Monitoring and Feedback Adjustment*. 1st edn. Wiley-Interscience.

[Box & Ramirez, 1992] Box, George, & Ramirez, Jose. 1992. Cumulative score charts. *Quality and Reliability Engineering International*, **8(1)**, 17–27.

[Boyens, 2004] Boyens, C., R. Krishnan R. Padman. 2004. On Privacy-Preserving Access to Distributed Heterogeneous Healthcare Information. *In: Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04).*

[Bradley *et al.*, 2005] Bradley, C. A., Rolka, H., Walker, D., & Loonsk, J. 2005. BioSense: Implementation of a National Early Event Detection and Situational Awareness System. *Morbidity and Mortality Weekly Report (MMWR)*, **54**, 11–19.

[Bravata *et al.*, 2004] Bravata, D.M., McDonald, M.M., Smith, W.M., Rydzak, C., Szeto, H., Buckeridge, D.L., Haberland, C., & Owens, D.K. 2004. Systematic Review: Surveillance Systems for Early Detection of Bioterrorism-Related Diseases. *Annals of Internal Medicine*, **140**, 910–922.

[Brillman *et al.*, 2005]Brillman, J. C., Burr, T., Forslund, D., Joyce, E., Picard, R., & Umland, E. 2005. Modeling emergency department visit patterns for infectious disease complaints: results and application to disease surveillance. *BMC Medical Informatics and Decision Making*, **5:4**, 1–14.

[Brockwell & Davis, 1987] Brockwell, P. J., & Davis, R. A. 1987. *Time series: theory and methods, 2nd ed.* Springer, New York.

[Brook & Evans, 1972] Brook, D., & Evans, D. A. 1972. An approach to the probability distribution of cusum run length. *Biometrika*, **59**, 539–549.

[Brookmeyer *et al.*, 2005] Brookmeyer, R., Johnson, E., & Barry, S. 2005. Modelling the incubation period of anthrax. *Statistics in Medicine*, **24(4)**, 531–542.

[Brookmeyer *et al.*, 2003] Brookmeyer, Ron, Johnson, Elizabeth, & Bollinger, Robert. 2003. *Modeling the optimum duration of antibiotic prophylaxis in an anthrax outbreak*. unpublished manuscript.

[Brown, 1959] Brown, R.G. 1959. *Statistical Forecasting for Inventory Control*. McGraw-Hill.

[Buckeridge *et al.*, 2005] Buckeridge, D. L., Burkom, H., Campbell, M., Hogan, W. R., & Moore, A. W. 2005. Algorithms for rapid outbreak detection: a research synthesis. *Journal of Biomedical Informatics*, **38**, 99–113.

[Buckeridge *et al.*, 2004] Buckeridge, David L., Burkom, H., Moore, A., Pavlin, J., Cutchis, P., & Hogan, W. 2004. Evaluation of Syndromic Surveillance Systems — Design of an Epidemic Simulation Model. *Morbidity and Mortality Weekly Report (MMWR)*, **53**, 137–143.

[Buckeridge *et al.*, 2008] Buckeridge, David L., Okhmatovskaia, Anna, Tu, Samson, O'Connor, Martin, Nyulas, Csongor, & Musen, Mark A. 2008. Predicting Outbreak Detection in Public Health Surveillance: Quantitative Analysis to Enable Evidence-Based Method Selection. *AMIA Annual Symposium proceedings*, **6**, 76–80.

[Buckeridge, 2007] Buckeridge, D.L. 2007. Outbreak. *Journal of Biomedical Informatics*, **40(4)**, 370–379.

[Buehler *et al.*, 2007] Buehler, James W., Isakov, Alexander P., Prietula, Michael J., Smith, Donna J., & Whitney, Ellen A. 2007. Preliminary Findings from the BioSense Evaluation Project. *Advances in Disease Surveillance*, **4**, 237.

[Burkom *et al.*, 2005] Burkom, H. S, and Murphy S., Coberly, J., & Hurt-Mullen, K. 2005. Public Health Monitoring Tools for Multiple Data Streams. *Morbidity and Mortality Weekly Report (MMWR)*, **54(Suppl)**, 55–62.

[Burkom & Murphy, 2007a] Burkom, H., & Murphy, S. 2007a. Data Classification for Selection of Temporal Alerting Methods for Biosurveillance. *Lecture Notes in Computer Science*, **4506**, 59.

[Burkom *et al.*, 2004] Burkom, H. S., Elbert, Y., Feldman, A., & Lin, J. 2004. Role of Data Aggregation in Biosurveillance Detection Strategies with Applications from ESSENCE. *Morbidity and Mortality Weekly Report (MMWR)*, **53**, 67–73.

[Burkom *et al.*, 2007] Burkom, H. S., Murphy, S. P., & Shmueli, G. 2007. Automated Time Series Forecasting for Biosurveillance. *Statistics in Medicine*, **26**, 4202–4218.

[Burkom, 2010] Burkom, Howard. 2010. Introducing the ISDS Biosurveillance Contest. *Advances in Disease Surveillance*, **forthcoming**.

[Burkom & Murphy, 2007b] Burkom, Howard, & Murphy, Sean. 2007b. Data Classification for Selection of Temporal Alerting Methods for Biosurveillance. *In: Second NSF Workshop, BioSurveillance 2007, New Brunswick, NJ, USA, May 22, 2007*.

[Burkom, 2003a] Burkom, Howard S. 2003a. Biosurveillance Applying Scan Statistics with Multiple, Disparate Data Sources. *Journal of Urban Health*, **80**, 57–65.

[Burkom, 2003b] Burkom, Howard S. 2003b. Development, Adaptation and Assessment of Alerting Algorithms for Biosurveillance. *Johns Hopkins APL Technical Digest*, **24**(4), 335–342.

[Burkom *et al.*, 2008] Burkom, Howard S., Loschen, Wayne A., Mnatsakanyan, Zaruhi R., & Lombardo, Joseph S. 2008. Tradeoffs Driving Policy and Research Decisions in Biosurveillance. *Johns Hopkins APL Tech. Digest*, **27(4)**, 299–312.

[Byrd *et al.*, 1995] Byrd, R.H., Lu, P., Nocedal, J., & Zhu, C. 1995. A limited memory algorithm for bound constrained optimization. *SIAM J. Scientific Computing*, **16**, 1190–1208.

[Casman, 2004] Casman, E. A. 2004. The Potential of Next-Generation Microbiological Diagnostics to improve Bioterrorism Detection Speed. *Risk Analysis*, **24**, 521–535.

[CDC, 2006] CDC. 2006. *CDC Syndromic Surveillance site*. http://www.cdc.gov/mmwr/pdf/wk/mm54su01.pdf.

[CDC, 2007] CDC. 2007. Norovirus Activity — United States, 2006–2007. *Morbidity and Mortality Weekly Report (MMWR)*, **56(33)**, 842–846.

[Chatfield, 1978] Chatfield, C. 1978. The Holt-Winters Forecasting Procedure. *Applied Statistics*, **27**, 264–279.

[Chatfield *et al.*, 2001] Chatfield, C., Koehler, A.B., Ord, J.K., & Snyder, R.D. 2001. A New Look at Models For Exponential Smoothing. *The Statistician, Journal of the Royal Statistical Society - Series D*, **50**(2), 147–159.

[Chatfield & Yar, 1991] Chatfield, Chris, & Yar, Mohammed. 1991. Prediction intervals for multiplicative Holt-Winters. *Internaional Journal of Forecasting*, **7**, 31–37.

[Cowling *et al.*, 2008] Cowling, Benjamin J, Lau, Eric H.Y., Lam, Conrad L.H., Cheng, Calvin K.Y., Kovar, Jana, Chan, Kwok Hung, Peiris, J.S. Malik, & Leung, Gabriel M. 2008. Effects of School Closures, 2008 Winter Influenza Season, Hong Kong. *Emerging infectious Diseases*, **14(10)**, 1660–1662.

[Crowder, 1987] Crowder, S. V. 1987. A simple method for studying run-length distributions of exponentially weighted moving average charts. *Technometrics*, **29**, 401–407.

[Czaplinski & Diers, 1998] Czaplinski, Cindy, & Diers, Donna. 1998. The effect of staff nursing on length of stay and mortality. *Medical Care*, **36(12)**, 1626–1638.

[Dafni *et al.*, 2004] Dafni, Urania G., Tsiodras, S., Panagiotakos, D., Gkolfinopoulou, K., Kouvatseas, G., Tsourti, Z., & Saroglou, G. 2004. Algorithm for Statistical Detection of Peaks — Syndromic Surveillance System for the Athens 2004 Olympic Games. *Morbidity and Mortality Weekly Report (MMWR)*, **53(Suppl)**, 86–94.

[Davis *et al.*, 2008] Davis, Mollie M., King Jr., James C., Moag, Lauren, Cummings, Ginny, & Magder, Laurence S. 2008. Countywide School-Based Influenza Immunization: Direct and Indirect Impact on Student Absenteeism. *Pediatrics*, **122(1)**, e260–e265.

[Dobra & Fienberg, 2001] Dobra, Adrian, & Fienberg, Stephen E. 2001. Bounds for cell entries in contingency tables induced by fixed marginal totals. *UNECE Statistical Journal*, **18**, 363–371.

[Dobra *et al.*, 2003] Dobra, Adrian, Fienberg, Stephen E., & Trottini, Mario. 2003. Assessing the Risk of Disclosure of Confidential Categorical Data. *In:* Bernardo, Jose M., Bayarri, M. J., Dawid, A. Philip, Berger, James O., Heckerman, D., Smith, A. F. M., & West, Mike (eds), *Bayesian Statistics 7, Proceedings of the Seventh Valencia International Meeting on Bayesian Statistics*. Oxford University Press.

[Doll & Hill, 1956] Doll, Richard, & Hill, Austin Bradford. 1956. Lung cancer and other causes of death in relation to smoking; a second report on the mortality of British doctors. *British Medical Journal*, **2 (5001)**, 1071–81.

[Domingo-Ferrer, 2002] Domingo-Ferrer, J., A. Oganian V. Torra. 2002. Information-Theoretic Disclosure Risk Measures in Statistical Disclosure Control of Tabular Data. *In: Proceedings of the 14th International Conference on Scientific and Statistical Database Management (SSDBM '02).*

[Duczmal *et al.*, 2006] Duczmal, Luiz, Kulldorff, Martin, & Huang, Lan. 2006. Evaluation of Spatial Scan Statistics for Irregularly Shaped Clusters. *Journal of Computational and Graphical Statistics*, **15(2)**, 428–442.

[Duncan *et al.*, 2001] Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R., & Roehrig, S. 2001. Disclosure limitation methods and information loss for tabular data. *Pages 135–166 of:* Doyle, P., Lane, J., Theeuwes, J., & Zayatz, L. (eds),

Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. Elsevier.

[Espino *et al.*, 2004] Espino, Jeremy U., Wagner, M., Szczepaniak, C., Tsui, F-C., Su, H., Olszewski, R., Liu, Z., Chapman, W., Zeng, X., Ma, L., Lu, Z., & Dara, J. 2004. Removing a Barrier to Computer-Based Outbreak and Disease Surveillance — The RODS Open Source Project. *Morbidity and Mortality Weekly Report (MMWR)*, **53**, 32–39.

[Farrington & Andrews, 2004] Farrington, C.P., & Andrews, N. 2004. Outbreak detection: application to infectious disease surveillance. *In:* Brookmeyer, R., & Stroup, D.F. (eds), *Monitoring the Health of Populations: Statistical Principles & Methods for Public Health Surveillance*. Oxford University Press.

[Farrington *et al.*, 1996] Farrington, C.P., Andrews, N.J., Beale, A.D., & Catchpole, M.A. 1996. A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **159**(3), 547–563.

[Fienberg, 2001] Fienberg, S. E. 2001. Statistical perspectives on confidentiality and data access in public health. *Statistics in Medicine*, **20**, 1347–1357.

[Fienberg & Shmueli, 2005] Fienberg, S. E., & Shmueli, G. 2005. Statistical Issues and Challenges Associated with Rapid Detection of Bio-terrorist Attacks. *Statistics in Medicine*, **24(4)**, 513–529.

[Fisher, 1922] Fisher, Ronald Aylmer. 1922. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society, A*, **222**, 309–368.

[Fricker & Rolka, 2006] Fricker, R.D., Jr., & Rolka, H.R. 2006. Protecting Against Biological Terrorism: Statistical Issues in Electronic Surveillance. *Chance*, **19**, 4–13.

[Fricker *et al.*, 2008a] Fricker, Ronald D., Jr., Knitt, Matthew C., & Hu, Cecilia X. 2008a. Comparing Directionally Sensitive MCUSUM and MEWMA Procedures with Application to Biosurveillance. *Quality Engineering*, **20(4)**, 478–494.

[Fricker, 2006] Fricker, Jr., R. D. 2006. Directionally Sensitive Multivariate Statistical Process Control Methods with Application to Syndromic Surveillance. *Advances in Disease Surveillance*, **3:1**, 1–17.

[Fricker *et al.*, 2008b] Fricker, Jr, Ronald D., Hegler, Benjamin L., & Dunfee, David A. 2008b. Comparing syndromic surveillance detection methods: EARS versus a CUSUM-based methodology. *Statistics in Medicine*, **27(17)**, 3407–29.

[Friedman & Rafsky, 1979] Friedman, Jerome H., & Rafsky, Lawrence C. 1979. Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests. *The Annals of Statistics*, **7(4)**, 697–717.

[Genz *et al.*, 2009] Genz, Alan, Bretz, Frank, Miwa, Tetsuhisa, Mi, Xuefei, Leisch, Friedrich, Scheipl, Fabian, & Hothorn, Torsten. 2009. *mvtnorm: Multivariate Normal and t Distributions*. R package version 0.9-7.

[Gesteland *et al.*, 2003] Gesteland, P.H., Gardner, R.M., Tsui, F.-C., Espino, J.U., Rolfs, R.T., James, B.C., Chapman, W.W., Moore, A.W., & Wagner, M.M.

2003. Automated Syndromic Surveillance for the 2002 Winter Olympics. *Journal of the American Medical Informatics Associaition*, **10**, 547–554.

[Ginsberg *et al.*, 2009] Ginsberg, Jeremy, Mohebbi, MatthewH., Patel, Rajan S., Brammer, Lynnette, Smolinski, Mark S., & Brilliant, Larry. 2009. Detecting influenza epidemics using search engine query data. *Nature*, **457**, 1012–1014.

[Goldenberg *et al.*, 2002a] Goldenberg, A., Shmueli, G., Caruana, R. A., & Fienberg, S. E. 2002a. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. *Proceeding of the National Academy of Sciences*, **99**, 5237–5240.

[Goldenberg *et al.*, 2002b] Goldenberg, A., Shmueli, G., & Caruana, R. A. 2002b. *Using grocery sales data for the detection of bio-terrorist attacks*. Unpublished manuscript.

[Grais *et al.*, 2007] Grais, R.F., Conlan, A.J.K., Ferrari, M.J, Djibo, A., Le Menach, A., Bjornstad, O.N., & Grenfell, B.T. 2007. Time is of the essence: exploring a measles outbreak response vaccination in Niamey, Niger. *Journal of the Royal Society Interface*, **5(18)**, 67–74.

[Green *et al.*, 2000] Green, Michael D., Freedman, D. Mical, & Gordis, Leon. 2000. Reference Manual on Epidemiology. *Pages 333–400 of: Reference Manual on Scientific Evidence, Second Edition*. LRP Publications.

[Guasticchi *et al.*, 2008] Guasticchi, G., Rossi, P. Giorgi, Lori, G., Genio, S., Biagetta, F., Gabriele, S., Pezzotti, P., & Borgia, P. 2008. Syndromic surveillance:

sensitivity and positive predictive value of the case definitions. *Epidemiology and Infection*, **21**, 1–10.

[Gwilym *et al.*, 2005] Gwilym, S., Howard, D.P.J., & Davies, N. 2005. Harry Potter casts a spell on accident prone children. *The British Medical Journal*, **331**, 1505 – 1506.

[Hall & Tajvidi, 2002] Hall, Peter, & Tajvidi, Nader. 2002. Permutation Tests for Equality of Distributions in High-Dimensional Settings. *Biometrika*, **89(2)**, 359–374.

[Heffernan *et al.*, 2004] Heffernan, R., Mostashari, F., Das, D., Besculides, M., Rodriguez, C., Greenko, J., Steiner-Sichel, L., Balter, S., Karpati, A., Thomas, P., Phillips, M., Ackelsberg, J., Lee, E., Leng, J., Hartman, J., Metzger, K., Rosselli, R., & Weiss, D. 2004. System Descriptions New York City Syndromic Surveillance Systems. *Morbidity and Mortality Weekly Report (MMWR)*, **53**, 23–27.

[Held *et al.*, 2005] Held, Leonhard, Hohle, Michael, & Hofmann, Mathias. 2005. A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, **5(3)**, 187–199.

[Henry & Fekete, 2006] Henry, Nathalie, & Fekete, Jean-Daniel. 2006. Evaluating visual table data understanding. *In: BELIV '06: Proceedings of the 2006 AVI workshop on Beyond time and errors: novel evaluation methods for information visualization.*

[Henze, 1988] Henze, Norbert. 1988. A Multivariate Two-Sample Test Based on the Number of Nearest Neighbor Type Coincidences. *The Annals of Statistics*, **16(2)**, 772–783.

[Hogan *et al.*, 2003] Hogan, William R., Tsui, Fu-Chiang, Ivanov, Oleg, Gesteland, Per H., Grannis, Shaun, Overhage, J. Marc, Robinson, J. Michael, & Wagner, Michael M. 2003. Detection of Pediatric Respiratory and Diarrheal Outbreaks from Sales of Over-the-counter Electrolyte Products. *Journal of the American Medical Informatics Association*, **10**(6), 555–562.

[Holt, 1957] Holt, C. C. 1957. *Forecasting seasonals and trends by exponentially weighted averages*. Tech. rept. Carnegie Institute of Technology.

[Hong & Hardin, 2005] Hong, Bo, & Hardin, J. Michael. 2005. A Study of the Performance of Multivariate Forecast-based Surveillance Schemes for Infectious Diseases on Multiple Locations. *In: Presentation at the 2005 Joint Statistical Meetings, Minneapolis, Minnesota*.

[Hope *et al.*, 2008a] Hope, K, Durrheim, DN, Muscatello, D, Merritt, T, Zheng, W, Massey, P, Cashman, P, & Eastwood, K. 2008a. Identifying pneumonia outbreaks of public health importance: can emergency department data assist in earlier identification? *Australian and New Zealand Journal of Public Health*, **32(4)**, 361–363.

[Hope *et al.*, 2008b] Hope, Kirsty, Merritt, Tony, Eastwood, Keith, Main, Kelly, Durrheim, David N, Muscatello, David, Todd, Kerry, & Zheng, Wei. 2008b. The public health value of emergency department syndromic surveillance following a natural disaster. *Communicable diseases intelligence*, **32(1)**, 92–4.

[Hutwagner *et al.*, 2003] Hutwagner, L., Thompson, W., Seeman, G.M., & Treadwell, T. 2003. The bioterrorism preparedness and response Early Aberration Reporting System (EARS). *Journal of Urban Health*, **80 (2) Suppl**, 89–96.

[Hutwagner *et al.*, 2005a] Hutwagner, L., Browne, T., Seeman, G. M., & Fleischauer, A. T. 2005a. Comparing Aberration Detection Methods with Simulated Data. *Emerging Infectious Diseases*, **11(2)**, 314–6.

[Hutwagner *et al.*, 2005b] Hutwagner, L. C., Thompson, W. W., Seeman, G. M., & Treadwell, T. 2005b. A simulation model for assessing aberration detection methods used in public health surveillance for systems with limited baselines. *Statistics in Medicine*, **24**(4), 543–550.

[Hutwagner *et al.*, 1997] Hutwagner, LC, Maloney, EK, Bean, NH, Slutsker, L., & Martin, SM. 1997. Using Laboratory-Based Surveillance Data for Prevention: An Algorithm for Detecting Salmonella Outbreaks. *Emerging Infectious Diseases*, **3**, 395–400.

[Ivanov *et al.*, 2003] Ivanov, O., Gesteland, P. H., Hogan, W., Mundorff, M. B., & Wagner, M. M. 2003. Detection of pediatric respiratory and gastrointestinal outbreaks from free-text chief complaints. *In: AMIA Annual Symposium*.

[Jaro, 1995] Jaro, M. A. 1995. Probabilistic linkage of large public health data files. *Statistics in Medicine*, **14**, 491–498. (disc: P687–689).

[Jobson, 1992]Jobson, J.D. 1992. *Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods*. Springer-Verlag, NY.

[Josseran *et al.*, 2009] Josseran, Loic, Caillere, Nadege, Brun-Ney, Dominique, Rottner, Jean, Filleul, Laurent, Brucker, Gilles, & Astagneau, Pascal. 2009.

Syndromic surveillance and heat wave morbidity: a pilot study based on emergency departments in France. *BMC Medical Informatics and Decision Making*, **9**.

[Jung *et al.*, 2006] Jung, Inkyung, Kulldorff, Martin, & Klassen, Ann. 2006. A Spatial Scan Statistic for Ordinal Data. *Statistics in Medicine*, **26(7)**, 1594 – 1607.

[Kaplan *et al.*, 2003] Kaplan, E.H., Patton, C.A., FitzGerald, W.P., & Wein, L.M. 2003. Detecting Bioterror Attacks by Screening Blood Donors: A Best-Case Analysis. *Emerging Infectious Diseases*, **9**, 909–914.

[Kikuchi *et al.*, 2007] Kikuchi, Kiyoshi, Ohkusa, Yasushi, Sugawara, Tamie, Taniguchi, Kiyosu, & Okabe, Nobuhiko. 2007. Syndromic Surveillance for Early Detection of Nosocomial Outbreaks. *Pages 202–208 of:* Zeng, Daniel, Gotham, Ivan, Komatsu, Ken, Lynch, Cecil, Thurmond, Mark, Madigan, David, Lober, Bill, Kvach, James, & Chen, Hsinchun (eds), *Intelligence and Security Informatics: Biosurveillance*. Springer Berlin / Heidelberg.

[Kim & Foutz, 1987] Kim, Kang-Kyun, & Foutz, Robert V. 1987. Tests for the Multivariate Two-Sample Problem Based on Empirical Probability Measures. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, **15(1)**, 41–51.

[Kleinberg, 2003] Kleinberg, Jon. 2003. Bursty and Hierarchical Structure in Streams. *Data Mining and Knowledge Discovery*, **7(4)**, 373–397.

[Kleinman & Abrams, 2006] Kleinman, K. P., & Abrams, A. M. 2006. Assessing surveillance using sensitivity, specificity and timeliness. *Statistical Methods in Medical Research*, **15(5)**, 445–464.

[Kleinman *et al.*, 2004] Kleinman, Ken, Lazarus, Ross, & Platt, Richard. 2004. A Generalized Linear Mixed Models Approach for Detecting Incident Clusters of Disease in Small Areas, with an Application to Biological Terrorism. *American Journal of Epidemiology*, **159(3)**, 217–224.

[Kovner & Gergen, 1998] Kovner, Christine, & Gergen, Peter J. 1998. Nurse staffing levels and adverse events following surgery in U.S. hospitals. *Journal of Nursing Scholarship*, **30(4)**, 315–321.

[Kulldorff, 1997] Kulldorff, Martin. 1997. A Spatial Scan Statistic. *Communications in Statistics–Theory and Methodology*, **26(6)**, 1481–1496.

[Kulldorff, 2001] Kulldorff, Martin. 2001. Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A*, **164**(1), 61–72.

[Kulldorff *et al.*, 2003] Kulldorff, Martin, Huang, Lan, & Pickle, Linda. 2003. An Elliptic Spatial Scan Statistic and Its Application to Breast Cancer Mortality Data in Northeastern United States. *Journal of Urban Health*, **80(suppl. 1)**, i130–i131.

[Lau *et al.*, 2008] Lau, Eric H. Y., Cowling, Benjamin J., Ho, Lai-Ming, & Leung, Gabriel M. 2008. Optimizing Use of Multistream Influenza Sentinel Surveillance Data. *Emerging Infectious Diseases*, **14(7)**, 1154–1157.

[Lawson, 2001] Lawson, A. 2001. Comments on the papers by Williams et al., Kulldorff, Knorr-Held and Best and Rogerson. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **164**(1), 97–99.

[Lazarus *et al.*, 2002] Lazarus, Ross, Kleinman, Ken, Dashevsky, Inna, Adams, Courtney, Kludt, Patricia, Alfred DeMaria, Jr., & Platt, Richard. 2002. Use of Automated Ambulatory-Care Encounter Records for Detection of Acute Illness Clusters, Including Potential Bioterrorism Events. *Emerging Infectious Diseases*, **8**, 753–760.

[Lee *et al.*, 2002] Lee, Ji-Eun, Pavlin, Julie, Elbert, Yevgeniy, & Kelley, Patrick. 2002. Analysis of a health indicator surveillance system: Its ability to detect annual influenza activity for the 1999-2000 and 2000-2001 seasons compared to traditional surveillance systems (presentation). *In: 2002 International Conference on Emerging Diseases*.

[Lombardo *et al.*, 2004] Lombardo, J. S., Burkom, H., & Pavlin, J. 2004. ESSENCE II and the Framework for Evaluating Syndromic Surveillance Systems. *Morbidity and Mortality Weekly Report (MMWR)*, **53(Suppl)**, 159–165.

[Loonsk, 2004] Loonsk, J.W. 2004. BioSense–a national initiative for early detection and quantification of public health emergencies. *Morbidity and Mortality Weekly Report (MMWR)*, **53**, 53–55.

[Lotze *et al.*, 2006] Lotze, Thomas, Shmueli, Galit, Murphy, Sean, & Burkom, Howard. 2006. A Wavelet-based Anomaly Detector for Early Detection of Disease Outbreaks. *In: Proceedings of the 23rd International Conference on Machine Learning (ICML), Workshop on Machine Learning Algorithms for Surveillance and Event Detection, Pittsburgh, PA*.

[Lotze *et al.*, 2008] Lotze, Thomas, Murphy, Sean P., & Shmueli, Galit. 2008. Preparing Biosurveillance Data for Classic Monitoring. *Advances in Disease Surveillance*, **6**, 1–20.

[Lotze & Shmueli, 2008a] Lotze, Thomas H., & Shmueli, Galit. 2008a. Ensemble Forecasting for Disease Outbreak Detection. *Pages 1470–1471 of: Proceedings of the 23rd AAAI Conference on Artficial Intelligence (AAAI-08), Chicago, IL.*

[Lotze & Shmueli, 2008b] Lotze, Thomas H., & Shmueli, Galit. 2008b. How does improved forecasting benefit detection? *International Journal of Forecasting*, **25(3)**, 467–483.

[Lotze *et al.*, 2010] Lotze, Thomas H., Shmueli, Galit, & Yahav, Inbal. 2010. Simulating and Evaluating Biosurveillance Datasets. *In:* Kass-Hout, Taha, & Zhang, Xaiohui (eds), *Biosurveillance: A Health Protection Priority (forthcoming; see http://www.routledgesociology.com/books/Biosurveillance-isbn9781439800461)*. Chapman and Hall.

[Lowen *et al.*, 2007] Lowen, Anice C., Mubareka, Samira, Steel, John, & Palese, Peter. 2007. Influenza Virus Transmission Is Dependent on Relative Humidity and Temperature. *Public Library of Science Pathogens*, **3(10)**.

[Maciejewski *et al.*, 2009] Maciejewski, Ross, Hafen, Ryan, Rudolph, Stephen, Tebbetts, George, Cleveland, William S., Ebert, David S., & Grannis, Shaun J. 2009. Generating synthetic syndromic-surveillance data for evaluating visual-analytics techniques. *IEEE Computer Graphics and Applications*, **29(3)**, 18–28.

[Mandl *et al.*, 2004] Mandl, KD, Reis, B, & Cassa, C. 2004. Measuring outbreak-detection performance by using controlled feature set simulations. *Morbidity and Mortality Weekly Report (MMWR)*, **53**, 130–136.

[Maragah & Woodall, 1992] Maragah, Hazem D., & Woodall, William H. 1992. The effect of autocorrelation on the retrospective X-chart. *Journal of Statistical Computation and Simulation*, **40**, 29–42.

[Marsden-Haug *et al.*, 2007] Marsden-Haug, N, Foster, V B, Gould, P L, Elbert, E, Wang, H, & Pavlin, J A. 2007. Code-based Syndromic Surveillance for Influenzalike Illness by International Classification of Diseases, Ninth Revision. *Emerging Infectious Diseases*, **13(2)**.

[Martinez-Beneito *et al.*, 2008] Martinez-Beneito, Miguel A., Conesa, David, Lopez-Quilez, Antonio, & Lopez-Maside, Aurora. 2008. Bayesian Markov switching models for the early detection of influenza epidemics. *Statistics in Medicine*, **27(22)**, 4455 – 4468.

[McCloskey, 1998] McCloskey, J M. 1998. Nurse staffing and patient outcomes. *Nursing Outlook*, **46(5)**, 199–200.

[Meselson *et al.*, 1994] Meselson, M., Guillemin, J., Hugh-Jones, M., Langmuir, A., Popova, I., Shelokov, A., & Yampolskaya, O. 1994. The Sverdlovsk anthrax outbreak of 1979. *Science*, **266(5188)**, 1202–1208.

[Meynard *et al.*, 2008] Meynard, Jean-Baptiste, Chaudet, HervÃ©, Texier, Gaetan, Ardillon, Vanessa, Ravachol, FranÃ§oise, Deparis, Xavier, Jefferson, Henry, Dussart, Philippe, Morvan, Jacques, & Boutin, Jean-Paul. 2008. Value of syndromic surveillance within the Armed Forces for early warning during a

dengue fever outbreak in French Guiana in 2006. *BMC Medical Informatics and Decision Making*, **8(1)**, 29(10).

[Monge & Elkan, 1996] Monge, Alvaro, & Elkan, Charles. 1996. The field-matching problem: Algorithm and application. *Pages 267–270 of: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.*

[Montgomery, 2001] Montgomery, D. C. 2001. *Introduction to Statistical Quality Control*. Third edn. John Wiley & Sons.

[Montgomery & Mastrangelo, 1991] Montgomery, D. C., & Mastrangelo, C. M. 1991. Some Statistical Process Control Methods for Autocorrelated Data. *Journal of Quality Technology*, **23**, 179–204.

[Moore *et al.*, 2002] Moore, Andrew, Cooper, Gregory, Tsui, Rich, & Wagner, Michael. 2002 (February). *Summary of Biosurveillance-relevant statistical and data mining technologies*. unpublished manuscript.

[Mostashari, 2002] Mostashari, F. 2002. *BT surveillance in NYC*. International Conference on Emerging Diseases (presentation).

[Mostashari *et al.*, 2003] Mostashari, F., Kulldorff, M., Hartman, J., Miller, J., & Kulasekera, V. 2003. Dead bird clusters as an early warning system for West Nile virus activity. *Emerging Infectious Diseases*, **9**, 641–646.

[Muscatello, 2004] Muscatello, D. 2004 (Novermber). *An adjusted cumulative sum for count data with day-of-week effects: application to influenza-like illness*. Presentation at 3rd National Syndromic Surveillance Conference.

[Najmi & Magruder, 2004] Najmi, A. H., & Magruder, S. F. 2004. Estimation of hospital emergency room data using otc pharmaceutical sales and least mean square filters. *BMC Medical Informatics and Decision Making*, **4**, 1–5.

[Najmi & Magruder, 2005] Najmi, A.H., & Magruder, S.F. 2005. An adaptive prediction and detection algorithm for multistream syndromic surveillance. *BMC Medical Informatics and Decision Making*, **12**, 5–33.

[Naus & Wallenstein, 2006] Naus, J., & Wallenstein, S. 2006. Temporal surveillance using scan statistics. *Statistics in Medicine*, **25(2)**, 311–324.

[Neill *et al.*, 2005] Neill, Daniel B., Moore, Andrew W., & Cooper, Gregory F. 2005. A Bayesian Spatial Scan Statistic. *Pages 1003–1010 of:* Weiss, Yair, Platt, John, & Scholkopf, Bernhard (eds), *Advances in Neural Information Processing Systems*, vol. 18.

[Neill *et al.*, 2007] Neill, Daniel B., Moore, Andrew W., & Cooper, Gregory F. 2007. A Multivariate Bayesian Scan Statistic. *Advances in Disease Surveillance*, **2:60**.

[NIST, 2004] NIST. 2004. *NIST/SEMATECH e-Handbook of Statistical Methods*. http://www.itl.nist.gov/div898/handbook/.

[Nobre & Stroup, 1994] Nobre, F.F., & Stroup, D.F. 1994. A Monitoring System to Detect Changes in Public Health Surveillance Data. *International Journal of Epidemiology*, **23**(2), 408–418.

[Noorossana & Vagjefi, 2005] Noorossana, R., & Vagjefi, S. J. M. 2005. Effect of Autocorrelation on Performance of the MCUSUM Control Chart. *Quality and Reliability Engineering International*, **22(2)**, 191–197.

[Ozonoff & Sebastiani, 2006] Ozonoff, A., & Sebastiani, P. 2006 (April). *Hidden Markov Models for Prospective Surveillance*. Presented at the Anomaly Detection group in National Defense and Homeland Security, SAMSI.

[Padgett *et al.*, 1992] Padgett, C. S., Thombs, L. A., & Padgett, W. J. 1992. On the alpha-Risks for Shewhart Control Charts. *Communications in Statistics-Simulation and Computation*, **21**, 1125–1147.

[Page, 1954] Page, E. S. 1954. Continuous inspection schemes. *Biometrika*, **41**, 100–115.

[Paladini, 2006] Paladini, M. 2006 (February). *From Data to Signals to Screenshots: Recent Developments in NYCDOHMH Emergency Department Syndromic Surveillance*. Presentation at DIMACS Working Group on BioSurveillance Data Monitoring and Information Exchange. Available at http://dimacs.rutgers.edu/Workshops/Surveillance/slides/paladini.ppt.

[Pavlin *et al.*, 2003] Pavlin, Julie A., Mostashari, Farzad, Kortepeter, Mark G., Hynes, Noreen A., Chotani, Rashid A., Mikol, Yves B., Ryan, Margaret A. K., Neville, James S., Gantz, Donald T., Writer, James V., Florance, Jared E., Culpepper, Randall C., Henretig, Fred M., & Kelley, Patrick W. 2003. Innovative Surveillance Methods for Rapid Detection of Disease Outbreaks and Bioterrorism: Results of an Interagency Workshop on Health Indicator Surveillance. *American Journal of Public Health*, **93**(8), 1230–1235.

[Polgreen *et al.*, 2008] Polgreen, P. M., Chen, Y., Pennock, D. M., & Forrest, N. D. 2008. Using internet searches for influenza surveillance. *Clinical Infectious Diseases*, **47**, 1443–1448.

[Que & Tsui, 2008] Que, J, & Tsui, F. C. 2008. A Multi-level Spatial Clustering Algorithm for Detection of Disease Outbreaks. *AMIA Annual Symposium proceedings*, **6**, 611–615.

[R Development Core Team, 2009] R Development Core Team. 2009. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Radaelli, 1992] Radaelli, Giovanni. 1992. Using the Cuscore technique in the surveillance of rare health events. *Journal of Applied Statistics*, **19(1)**(1), 75–81.

[Reingold, 2003] Reingold, A. 2003. If Syndromic Surveillance is the Answer, What is the Question? *Biosecurity and Bioterrorism: Biodefense Strategy, Practice and Science*, **1**(2), 1–5.

[Reinke, 1991]Reinke, William A. 1991. Applicability of Industrial Sampling Techniques to Epidemiologic Investigations: Examination of an Underutilized Resource. *American Journal of Epidemiology*.

[Reis *et al.*, 2003] Reis, Ben Y., Pagano, Marcello, & Mandl, Kenneth D. 2003. Using temporal context to improve biosurveillance. *Proceedings of the National Academy of Sciences*, **100(4)**, 1961–1965.

[Reis & Mandl, 2003] Reis, B.Y., & Mandl, K.D. 2003. Time series modeling for syndromic surveillance. *BMC Medical Informatics and Decision Making*, **3**(2), 858–64.

[Rencher, 2002] Rencher, Alvin C. 2002. *Methods of Multivariate Analysis*. Wiley-Interscience.

[Rice, 1995] Rice, J. A. 1995. *Mathematical Statistics and Data Analysis, Second Edition*. Duxbury Press.

[Riffenburgh & Cummins, 2006] Riffenburgh, R.H., & Cummins, K.M. 2006. A simple and general change-point identifier. *Statistics in Medicine*, **25(6)**, 1067–1077.

[Rogerson & Yamada, 2004] Rogerson, Peter A., & Yamada, I. 2004. Approaches to syndromic surveillance when data consist of small regional counts. *Morbidity and Mortality Weekly Report (MMWR)*, **53**, 79–85.

[Rolka, 2006] Rolka, H. 2006. Emerging Public Health Biosurveillance Directions. *Pages 101–107 of:* Wilson, A, Wilson, G, & Olwell, D H (eds), *Statistical Methods in Counter-Terrorism: Game Theory, Modeling, Syndromic Surveillance and Biometric Authentication*. Springer.

[Schilling, 1986] Schilling, Mark F. 1986. Multivariate Two-Sample Tests Based on Nearest Neighbors. *Journal of the American Statistical Association*, **81(395)**, 799–806.

[Serfling, 1963] Serfling, R. E. 1963. Methods for current statistical analysis for excess pneumonia-influenza deaths. *Public Health Reports*, **78**, 494–506.

[Shmueli, 2005] Shmueli, G. 2005. *Wavelet-Based Monitoring for Modern Biosurveillance*. Tech. rept. RHS-06-002, University of Maryland, Robert H Smith School of Business.

[Shmueli & Fienberg, 2006] Shmueli, G., & Fienberg, S. E. 2006. Current and Potential Statistical Methods for Monitoring Multiple Data Streams for Bio-Surveillance. *Pages 109–140 of:* A Wilson, G Wilson, & Olwell, D H (eds),

Statistical Methods in Counter-Terrorism: Game Theory, Modeling, Syndromic Surveillance and Biometric Authentication. Springer.

[Shmueli & Burkom, 2009] Shmueli, Galit, & Burkom, Howard S. 2009. Statistical Challenges in Modern Biosurveillance. *Technometrics (Special Issue on Anomaly Detection)*, **forthcoming**.

[Shmueli *et al.*, 2005] Shmueli, Galit, Minka, Thomas P., Kadane, Joseph B., Borle, Sharad, & Boatwright, Peter. 2005. A Useful Distribution for Fitting Discrete Data: Revival of the COM-Poisson. *Journal of the Royal Statistical Society C*, **54**(1), 127–142.

[Shmueli *et al.*, 2007] Shmueli, Galit, Lotze, Thomas, & Yahav, Inbal. 2007. *Simulating Multivariate Syndromic Time Series and Outbreak Signatures*. Tech. rept. University of Maryland, Robert H. Smith School.

[Shtatland *et al.*, 2009] Shtatland, Ernest S., Kleinman, Ken, & Cain, Emily M. 2009. Biosurveillance and Outbreak Detection Using the ARIMA and LOGISTIC Procedures. *In: SAS SUGI 31 proceedings: Statistics, Data Analysis and Data Mining*.

[Shu *et al.*, 2007] Shu, Lianjie, Jiang, Wei, & Wu, Shujin. 2007. A One-Sided EWMA Control Chart for Monitoring Process Means. *Communications in Statistics - Simulation and Computation*, **36:4**, 901–920.

[Siddiqi *et al.*, 2007] Siddiqi, Sajid M., Boots, Byron, Gordon, Geoffrey J., & Dubrawski, Artur W. 2007. Learning Stable Multivariate Baseline Models for Outbreak Detection. *Advances in Disease Surveillance*, **4**, 266.

[Siegmund, 1985] Siegmund, D. 1985. *Sequential Analysis: Tests and Confidence Intervals*. Springer-Verlag.

[Siegrist & Pavlin, 2004] Siegrist, D., & Pavlin, J. 2004. Bio-ALIRT Biosurveillance Detection Algorithm Evaluation. *Morbidity and Mortality Weekly Report (MMWR)*, **53**, 152–158.

[Siegrist *et al.*, 2005] Siegrist, D., McClellan, G., Campbell, M., Foster, V., Burkom, H., Hogan, W., Cheng, K., Buckeridge, D., Pavlin, J., & Kress, A. 2005. *Evaluation of Algorithms for Outbreak Detection Using Clinical Data from Five U.S. Cities*. Tech. rept. DARPA Bio-ALIRT Program.

[Sokolow *et al.*, 2005]Sokolow, Leslie Z., Grady, N., Rolka, H., Walker, D., McMurray, P., English-Bullard, R, & Loonsk, J. 2005. Deciphering data anomalies in BioSense. *Morbidity and Mortality Weekly Report (MMWR)*, **53**, 133–139.

[Sonesson & Bock, 2003] Sonesson, C., & Bock, D. 2003. A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **166**(1), 5–21.

[Stacey *et al.*, 2005] Stacey, D., Calvert, D., Shu, J., & Harvey, N. 2005. *A Preliminary Wavelet Analysis of OTC Pharmaceutical Sales Data*. Tech. rept. University of Guelph.

[Stoto *et al.*, 2006] Stoto, M., Fricker, R. D., Jain, A., Davies-Cole, J. O., Glymph, C., Kidane, G., Lum, G., Jones, L., Dehan, K., & Yuan, C. 2006. Evaluating Statistical Methods for Syndromic Surveillance. *Pages 141–172 of:* Wilson, A, Wilson, G, & Olwell, D H (eds), *Statistical Methods in Counter-Terrorism: Game*

Theory, Modeling, Syndromic Surveillance and Biometric Authentication. ASA-SIAM.

[Stoto *et al.*, 2004] Stoto, M.A., Schonlau, M., & Mariano, L.T. 2004. Syndromic Surveillance: Is it Worth the Effort? *Chance*, **17**(1), 19–24.

[Strzalka & Havens, 1996] Strzalka, Amy, & Havens, Donna S. 1996. Nursing care quality: comparison of unit-hired, hospital float pool, and agency nurses. *Journal of Nursing Care Quality*, **10(4)**, 59–65.

[Sullivan, 2003] Sullivan, B. M. 2003. Bioterrorism Detection: The Smoke Alarm and the Canary. *Technology Review Journal*, **11**(1), 135–140.

[Tarnow-Mordi *et al.*, 2000] Tarnow-Mordi, W. O., Hau, C., Warden, A., & J., Shearer A. 2000. Hospital mortality in relation to staff workload: a 4-year study in an adult intensive-care unit. *The Lancet*, **356(9225)**, 185–189.

[Tsui *et al.*, 2003] Tsui, Fu-Chiang, Espinoso, Jeremy U., Dato, Virginia M., Gesteland, Per H., Hutman, Judith, & Wagner, Michael M. 2003. Technical Description of RODS: A Real-time Public Health Surveillance System. *Journal of the American Medical Informatics Association*, **10(5)**, 399–408.

[Van Dobben De Bruyn, 1967] Van Dobben De Bruyn, C. S. 1967. The Interplay of Tracking Signals and Adaptive Predictors. *The Statistician*, **17(3)**, 237–246.

[Wagner *et al.*, 2001] 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., & Fridsma, D. B. 2001. The Emerging science of very early detection of disease outbreaks. *Journal of Public Health Management and Practice*, **7**, 51–59.

[Wagner *et al.*, 2003] Wagner, M W, Robinson, J M, Tsui, F-C, Espino, J U, & Hogan, W R. 2003. Design of a National Retail Data Monitor for Public Health. *Journal of the American Medical Informatics Association*, **10**(5), 409–418.

[Wagner *et al.*, 2006] Wagner, Michael M., Pavlin, Julie, Cox, Kenneth L., & Cirino, Nick M. 2006. Other Organizations That Conduct Biosurveillance. *Pages 183–196 of:* Wagner, Michael M., & Moore, Andrew W. (eds), *Handbook of Biosurveillance*. Elsevier.

[Wallenstein & Naus, 2004] Wallenstein, S., & Naus, J. 2004. Scan Statistics for Temporal Surveillance for Biologic Terrorism. *Morbidity and Mortality Weekly Report (MMWR)*, **53(Suppl)**, 74–78.

[Wallstrom *et al.*, 2005] Wallstrom, Garrick L., Wagner, M., & Hogan, W. 2005. High-fidelity injection detectability experiments: a tool for evaluating syndromic surveillance systems. *Morbidity and Mortality Weekly Report (MMWR)*, **54**, 85–91.

[Watkins *et al.*, 2007] Watkins, Rochelle E, Eagleson, Serryn, Beckett, Sam, Garner, Graeme, Veenendaal, Bert, Wright, Graeme, , & Plant, Aileen J. 2007. Using GIS to create synthetic disease outbreaks. *BMC Medical Informatics and Decision Making*, **7:4**.

[Wheeler, 1991] Wheeler, D.J. 1991. Shewhart's Charts: Myths, Facts and Competitors. *In: ASQC Quality Congress Transactions*. Milwaukee, WI.

[Wheeler, 1992] Wheeler, Donald J. 1992. Correlated Data and Control Charts. *In: Fifth Annual Forum of the British Deming Association*.

[Widdowson *et al.*, 2003] Widdowson, M-A., Bosman, A., van Straten, E., Tinga, M., Chaves, S., van Eerden, L., & van Pelt, W. 2003. Automated, Laboratory-based System Using the Internet for Disease Outbreak Detection, the Netherlands. *Emerging Infectious Diseases*, **9(9)**, 1046–1052.

[Wilkening, 2008] Wilkening, Dean A. 2008. Modeling the Incubation Period of Inhalational Anthrax. *Medical Decision Making*, **28**, 593–606.

[Wong *et al.*, 2003a] Wong, W.-K., Moore, A., Cooper, G., & Wagner, M. 2003a. Bayesian Network Anomaly Pattern Detection for Disease Outbreaks. *Pages 808–815 of: Proceedings of the Twentieth International Conference on Machine Learning*. Menlo Park, California: AAAI Press.

[Wong, 2004] Wong, Weng-Keen. 2004. *Data Mining for Early Disease Outbreak Detection*. Ph.D. thesis, School of Computer Science, Carnegie Mellon University.

[Wong *et al.*, 2002] Wong, Weng-Keen, Moore, Andrew, Cooper, Gregory, & Wagner, Michael. 2002. Rule-based Anomaly Pattern Detection for Detecting Disease Outbreaks. *In: Proceedings of the 18th National Conference on Artificial Intelligence*. MIT Press.

[Wong *et al.*, 2003b] Wong, Weng-Keen, Moore, Andrew, Cooper, Gregory, & Wagner, Michael. 2003b. What's Strange About Recent Events. *Journal of Urban Health*, **80**(June), i66–i75.

[Wong *et al.*, 2005] Wong, W.K., Moore, A., Cooper, G., & Wagner, M. 2005. What's Strange About Recent Events (WSARE): An Algorithm for the Early

Detection of Disease Outbreaks. *The Journal of Machine Learning Research*, **6**, 1961–1998.

[Woodall, 2006] Woodall, W. H. 2006. The Use of Control Charts in Health-Care and Public-Health Surveillance. *Journal of Quality Technology*, **38(2)**, 89–104.

[Woodall & Faltin, 1993] Woodall, W. H., & Faltin, F. W. 1993. Autocorrelated Data and SPC. *ASQC Statistics Division Newsletter*, **13**, 18–21.

[Yahav & Shmueli, 2007] Yahav, Inbal, & Shmueli, Galit. 2007. *Evaluating Directionally-Sensitive Multivariate Control Charts with an Application to Biosurveillance*. Tech. rept. RHS-06-059, Robert H Smith School, University of Maryland (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1119279).

[Zeileis *et al.*, 2009] Zeileis, A., Hornik, K., & Murrell, P. 2009. Escaping RGBland: Selecting Colors for Statistical Graphics. *Computational Statistics & Data Analysis*, **53**, 3259–3270.

[Zhang *et al.*, 2003] Zhang, J., Tsui, F.C., Wagner, M.M., & Hogan, W.R. 2003. Detection of outbreaks from time series data using wavelet transform. *Pages 748–752 of: AMIA Annual Symposium Proceedings*.

[Zhang *et al.*, 2008] Zhang, Min, Kong, Xiaohui, & Wallstrom, Garrick L. 2008. Simulation of Multivariate Spatial-Temporal Outbreak Data for Detection Algorithm Evaluation. *Pages 155–163 of: Biosurveillance and Biosecurity*. Springer.

[Zheng *et al.*, 2007] Zheng, Wei, Aitken, Robert, Muscatello, David J, & Churches, Tim. 2007. Potential for early warning of viral influenza activity in the

community by monitoring clinical diagnoses of influenza in hospital emergency departments. *BMC Public Health*, **7:250**.