12-2012

Using Box-Jenkins Modeling Techniques to Forecast Future Disease Burden and Identify Disease Aberrations in Public Health Surveillance Report

Larry C. Garrett
Western Michigan University, lgarrett@healthinsight.org

Recommended Citation

Garrett, Larry C., "Using Box-Jenkins Modeling Techniques to Forecast Future Disease Burden and Identify Disease Aberrations in Public Health Surveillance Report" (2012). *Dissertations*. 106.
https://scholarworks.wmich.edu/dissertations/106

USING BOX-JENKINS MODELING TECHNIQUES TO FORECAST FUTURE DISEASE BURDEN AND IDENTIFY DISEASE ABERRATIONS IN PUBLIC HEALTH SURVEILLANCE REPORT

by

Larry C. Garrett

A Dissertation
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Philosophy
Interdisciplinary Health Sciences
Advisor: Kieran Fogarty, Ph.D.

Western Michigan University
Kalamazoo, Michigan
December 2012

USING BOX-JENKINS MODELING TECHNIQUES TO FORECAST FUTURE DISEASE BURDEN AND IDENTIFY DISEASE ABERRATIONS IN PUBLIC HEALTH SURVEILLANCE REPORT

Larry C. Garrett, Ph.D.

Western Michigan University, 2012

The analysis of public health surveillance data to identify departures from historical patterns of disease is required to facilitate the timely identification of potential outbreaks. Using the Box-Jenkins forecasting model, this study examines the potential to predict future disease burden based upon the historical record within local public health jurisdictions. Box-Jenkins forecasting was developed as a direct result of forecast problems in the business, economic, and control-engineering applications, yet it has not been systematically examined for use with public heath surveillance data.

Box-Jenkins forecast models are constructed by stratifying 84,029 disease reports from the State of Utah by year (*n* = 10), disease type (*n* = 50), and jurisdiction (*n* = 13). A disease has to be present in all years and have a rate greater than 0.2/100K to be included in the study. Sixteen diseases have been selected for analysis. Accuracy of the forecasts is determined by conducting 48 forecast trials; within these trials there are 576 monthly forecasts. The results are compared to the actual values for the same period. Accuracy is determined calculating the Mean Absolute Percentage Error (MAPE) for each forecast trial.

Forecast predication intervals explore the relationship between actual values and the predication interval associated with each forecast.

Forecasts have an absolute accuracy of 71% (range: 43.4–91.7%). Ten of the 16 forecasts (63%) have an absolute accuracy greater than 75%, four (25%) have an absolute accuracy between 52.6% and 69.6%, and two (12%) have an accuracy of less than 50%. Forecast accuracy is independent of rate of disease ($r = -.348$, $n = 16$, $p > .05$) and jurisdictional size ($r = .396$, $n = 7$, $p = .380$). Eighty-four percent of all forecast values are contained within the first forecast interval, 88% within the second, and 99% within the third.

This study demonstrates that it is possible to predict future disease burden using Box-Jenkins forecasting techniques. The overall accuracy of the forecast and disproportionate number of forecast values contained within the first forecast interval validate this as a method that may be used to monitor disease trends and potentially facilitate the early identification of an outbreak.

ACKNOWLEDGMENTS

I would like to acknowledge Dr. Kieran Fogarty and Dr. Amy Curtis for their guidance and support as I undertook my studies in the Interdisciplinary Health Sciences Ph.D. program at Western Michigan University, College of Health and Human Services. Their encouragement and guidance throughout my doctoral studies is genuinely appreciated. I also want to thank Dr. George White at Westminster College and Dr. Robert Walsh and Utah Valley University for their frank review of my work and ability to provide constructive criticism and guidance in a positive and clear manner.

Additionally, I would like to thank my cohort members: Dawn, Barbara, Stephanie, Amy, Theresa, Dae, Jody, Patricia, and Melisa; thanks for your friendship and for the study times we shared.

Finally, I would like to thank my wife, Deborah, and my children, Curtis, James, and Erika for their patience and genuine encouragement over these past years.

Larry C. Garrett

TABLE OF CONTENTS

CHAPTER

Table of Contents—Continued

CHAPTER

APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The analysis of public health surveillance data to identify departures from previously observed or historical patterns of disease is required to facilitate the timely identification of a potential outbreak or epidemic. It is believed by many that if an outbreak is detected early the public health response may limit its spread and break the chain of transmission thereby reducing the overall disease of the outbreak burden on a community (Williamson & Weatherby, 1999).

For definitional purposes, an epidemic (from the Greek *epi* [upon], *demos* [people]) (Berube, 1985) occurs when new cases of a disease, in a given population, and during a defined period, substantially exceeds what is expected (Last, 1995). As such, baselines or expected counts from the historical record of disease events are needed to determine whether a disease report represents an expected event or may be part of a larger problem or epidemic. For example, a high count of influenza cases during the winter in the northern latitudes is expected, whereas the same number of cases during the summer may be cause for concern (Lipsitch & Viboud, 2009).

To create comparative baselines, epidemiologists use data from disease surveillance systems with disease report information obtained from physicians, hospitals, public health departments, and public and private laboratories (Teutsch & Churchill, 1994). To be relevant, these baselines need to be created on a continual basis and grounded upon sound forecasting methodologies. Presently, many of the baselines used in public health rely upon basic methods such as historical counts of disease (Overhage, Grannis, & McDonald, 2008) or forecasts created using simple moving averages (SMA). The SMA is the most basic forecasting method and tends to produce good forecasts only if the data are relatively stable or slowly changing as volatile and unstable data elements produce unsatisfactory results (Makridakis, Wheelwright, & Hyndman, 2000). However, many infectious disease patterns are not stable due to increasing or decreasing populations, seasonality of disease, the occurrence of outbreaks and other external conditions (Unkel, Farrington, Garthwaite, Robertson, & Andrews, 2012).

Statement of Problem

Due in part to the events surrounding 9/11, there has been a renewed interest in public health disease surveillance systems and the methods used to analyze their data (Gesteland et al., 2002). Additional concerns that have also driven this interest include emerging infectious diseases (e.g., Severe Acute Respiratory Syndrome), bioterrorism (e.g., Anthrax), and concerns over globalization and the subsequent transmission of disease (e.g., Influenza) (Mikanatha, 2007). Much of this interest has focused upon new surveillance methodologies that are collectively known as syndromic surveillance systems (Buckeridge, 2010). These systems, used by public health practitioners, are intended to give advanced notification of a potential outbreak. They are based upon syndromes, symptoms of disease such as a fever and rash, or data from secondary sources such as the amount of antidiarrheal medication sold at a local pharmacy. New methods, including the establishment of moving averages and baselines, are necessary for the analysis of these data. However, the analytical methodology associated with traditional disease report surveillance systems based upon actual diagnosed disease reports has remained relatively unchanged during this period. With or without these new systems, public health professionals need to be capable of monitoring disease trends and accessing information from multiple sources to identify or characterize situations that may signal an outbreak or other public health emergency (Goldstein, 2010). This dissertation focuses strictly upon the analysis of data obtained from public health surveillance systems compatible with the National Electronic Disease Surveillance System (NEDSS). NEDSS compatible systems are based upon actual disease reports originating from physicians, hospitals and laboratories and not upon syndromes or other secondary data sources. Data from a NEDSS system will be examined to determine if accurate disease forecasts can be made using Box-Jenkins forecasting models. Box-Jenkins was selected as many diseases exhibit trend and seasonality (Uziel & Stone, 2012). This model takes into account such occurrences when constructing forecasts (Box & Jenkins, 1994). As such, Box-Jenkins may be an appropriate tool to make disease forecasts; it has not been systematically evaluated for use with NEDSS data.

Definition of Terms

The vocabulary associated with forecasting is often unclear due to the assortment of terms used to describe the outcome or purpose of a forecast. This may lead to confusion and it is therefore necessary to define these terms to clarify the intent of the forecasting processes associated with this dissertation. Terms frequently associated with forecasting include: (a) forecast, (b) prediction, (c) scenario, (d) extrapolation, and (e) projection. The following quote provides the context used to define these terms.

> A *forecast* is a probabilistic statement, on a relatively high confidence level, about the future. A *prediction* is an apodictic (non-probabilistic) statement, on an absolute confidence level, about the future. An *anticipation* is a logical constructed model of a possible future, on a confidence level as yet undefined. (Jantch, 1967)

Using the above quote as a framework, a definition of terms associated with forecasting follows:

- A statement about the future that has a quantifiable probability of being accurate is a forecast.
- A statement about the future put forward as a certainty but not based upon statistical modeling, and often an opinion, is a prediction.
- A statement of a possible future without a quantified probability is a scenario.
- A statement about the future based upon the continuation of a past trend is an extrapolation.

- A statement based upon a scenario or extrapolation and contains a far-reaching view of the future with consideration of past and current events is a projection.

For consistency, the forecast discussion presented in this dissertation is based upon statistical modeling and outcomes that are based upon a quantifiable probability. As such, forecasting, in its simplest terms, is a systematic process with the objective of predicting and making statements about events whose actual outcomes have not yet been observed. More precisely, forecasting attempts to predict change in the presence of uncertainty. This uncertainty is based upon trend, cycle, and seasonality (Levenback & Cleary, 2006). Forecasting uses recognized statistical methods utilizing historical data from longitudinal data sets (Armstrong, 2001). There is no single correct forecasting method to use and the method selection is usually based on the objectives associated with the forecast and the underlying condition of the data used to create the forecast. Forecasting methods can be broken into time series and explanatory types of analysis (Makridakis & Wheelwright, 1987). Time series models lend themselves to predicting the continuation of historical patterns, such as disease burden within a community, if the three following conditions are met:

1. Data from the historical record are available.
2. These data are quantified in the form of numerical data.
3. It can be assumed that at least some portion of the past pattern will continue into the future.

The last condition, known as the assumption of continuity, is the underlying premise of all quantitative forecasting methods (Makridakis et al., 2000).

Explanatory forecasting seeks to understand an action or phenomena such as exploring how weather patterns affect asthma rates within a community. Qualitative forecasts are used when there are little or no quantitative data or when one believes they have unique insight into the future. Research has consistently shown that the judgment of humans is usually less accurate than those associated with even simple quantitative models (Hogarth & Makridakis, 1981); as such, they are not discussed further.

Forecast accuracy is the difference between the forecast and actual value during a defined time period (Armstrong & Collopy, 1992). Just as there is no single correct forecasting method, the determination of the accuracy of a forecast varies depending upon the objective of the forecast. The method to ascertain the accuracy in this analysis is the mean absolute percentage error (MAPE). MAPE compares the forecast values against the actual values for each forecast trial. Additional details about MAPE and the methods used to determine forecast accuracy are described in detail in the Methods chapter.

Significance of Research

Box-Jenkins was selected for evaluation because it has the potential of producing a point forecast within a given population, it provides a forecast interval, and is based upon a proven model (Geurts & Ibrahim, 1975); however, it has not been systematically evaluated for use in public health. Forecast results and their associated forecast intervals may help local and state public health practitioners make informed decisions about whether the number of observed disease reports in a given timeframe represents a potential outbreak or is a function of random variation. This is possible as Box-Jenkins relies upon the mathematical properties of the underlying time series from which the forecast is based and not upon the dynamics of infectious disease transmission.

Box-Jenkins is an autoregressive integrated moving average (ARIMA) model. The difference between traditional regression and ARIMA is that the variable being forecast is not related to another variable but is related to its own past values, a process known as autocorrelation (Levenback & Cleary, 2006). Autocorrelation examines the correlation between each observation and its previous observations. Moreover, forecasts based upon ARIMA take into account the premise that data, taken over time, may have an internal structure based upon trend, and seasonality that can be accounted for (Box & Jenkins, 1994). Many diseases exhibit trend and seasonality (Uziel & Stone, 2012), and as such, the use of Box-Jenkins is an appropriate tool to make these forecasts; it has not been systematically evaluated for use with NEDSS data.

Box-Jenkins is considered by many to be complex, which may help explain its limited use in public health and other disciplines including those that rely upon forecasting in the manufacturing, financial, and marketing industries (Levenback & Cleary, 2006; Mentzer & Cox, 1984; Winklhofer, Diamantopoulos, & Witt, 1996). Its usage may increase due to the number of statistical programs that support it use including SAS and SPSS; however, many of these programs require a certain level of expertise on the part of the user to create forecasts. This situation may be changing as there are an increasing number of computer programs that specialize in forecasting. Generally, these programs contain limited overall functionality when compared to complete statistical programs but offer an array of forecasting models and provide simplified methods for data input, analysis, and subsequent evaluation of the forecast models.

Purpose of the Study

This study examines the potential to predict future disease burden based upon the historical record within public health jurisdictions using Box-Jenkins forecasting models. It also examines the influence of jurisdictional size and rate of disease upon the accuracy of a forecast. The results are important as they may help facilitate the future identification of outbreaks and other disease related events. Moreover, these results will support the development of thresholds for proper disease forecasting methodologies and its appropriate use based upon jurisdictional size and rate of disease.

Research Questions

There are a limited number of articles describing disease forecasting using the Box-Jenkins method. However, the literature is silent on methods to determine disease forecast accuracy; as such, the threshold upon which a disease forecast is determined to be accurate was the subject of extensive consideration. Setting the threshold too low dilutes a model’s value, while setting it too high is impracticable as prophetic forecasting models do not exist. For this analysis, the threshold was determined by one-on-one interviews with state and local public health officials as well as individuals from academic settings. Participants include: three local health officers, three local health department epidemiologists, one disease investigator, one state epidemiologist, and two associate professors from schools of public health. To prepare for the interviews, participants were provided with the results of a pilot study upon which this dissertation is based and asked to consider the following question:

> If your surveillance data indicated that the number of hepatitis A disease reports exceeded the forecast baseline, how confident would you have to be with your baseline before you would consider publicly naming a suspect restaurant?

Hepatitis A was purposefully selected as the disease for these discussions based upon the following criteria:

- There is an effective public health intervention based upon the use of Immunoglobulin (IG) yet this intervention is time sensitive and potentially expensive based upon the number of necessary injections (Heymann, 2008).
- A public announcement of a hepatitis A associated with a restaurant carries financial implications. Food safety managers cite losses of 40% to 80% of revenues for a named restaurant (“Restaurant Industry,” 1997). An announcement that is later shown to be unwarranted may have legal ramifications.

- It is expensive for health departments to investigate an outbreak of hepatitis A associated with a restaurant. For example, the calculated cost associated with an outbreak in Denver was $689,314 (Dalton, Haddix, Hoffman, & Mast, 1996). An investigation that yields no results is costly in both time and other investigation related expenses.

The consensus of those interviewed was that they needed to know that a forecast was between 70–80 % accurate before they would consider publicly naming a restaurant. Based upon this, for this analysis, a forecast is considered accurate at the 75% threshold level.

Based upon the 75% accuracy threshold, two questions are answered in this dissertation, with the second question based upon the results of the first. Specifically, the research questions are:

1. Can Box-Jenkins forecasts produce disease specific forecasts that are equal to or greater than 75% accurate?
2. For diseases specific forecasts that are equal to or greater than 75% accurate, what influence does jurisdictional size and rate of disease have upon the accuracy of the forecasts?

The data source, diseases selected for evaluation, and the methods used to define forecast accuracy and to assess what influence jurisdictional size and rate of disease has upon these forecasts is described in the methods section of this dissertation.

Summary

Data within public health systems are used to help identify potential outbreaks or epidemics. To maximize the benefit of these data it is necessary to create baselines or expected counts. These baselines need to be created on a continual basis and grounded upon sound forecasting methodologies in order to give the forecast credibility.

This dissertation examines the use of Box-Jenkins based forecasting as a tool for public health use. It was selected for evaluation as it has the potential of producing accurate forecasts of disease within a given population as well as providing a forecast interval for a given forecast. Box-Jenkins is based upon an autoregressive integrated moving average model and is used in a variety of other professional disciplines that rely upon forecasting to help make decisions based upon data.

To determine the accuracy of the forecasts associated with this dissertation it is necessary to calculate the mean absolute percentage error for each forecast. Disease forecasts that are determined to be accurate (i.e., > 75% accurate) will be examined to determine at which point the Box-Jenkins methodology fails based upon rate of disease and population size.

The results of these analyses, will be used to aid in determining if the forecasting methodologies associated with Box-Jenkins can help local and state public health practitioners make informed decisions about whether the number of observed disease reports in a given timeframe represents a potential outbreak or is a function of random variation

CHAPTER II

LITERATURE REVIEW

To establish a contextual framework for this dissertation a comprehensive literature review was completed. This chapter presents the results of the review and includes: (a) the literature scope and search strategy, (b) the development of public health surveillance systems, (c) current methods and strategies used to analyze surveillance data based upon the National Notifiable Diseases Surveillance System (NNDSS[1]) architecture, (d) a description of Box-Jenkins, and (e) methods used to evaluate forecasting results. The intent is to present the current state of public health surveillance and analytical methodologies which in turn influences the overall scheme of this dissertation.

Literature Scope and Search Strategy

To assure completeness, an examination was completed on a wide array of literature that describes public health surveillance systems and the analytical methods used within these systems. Journal articles were reviewed describing the use of Box-Jenkins in forecasting, regardless of their professional affiliation

---

[1] The National Notifiable Diseases Surveillance System (NNDSS) represents public health surveillance systems developed by local and state-based public health agencies and the Centers for Disease Control and Prevention. These systems support surveillance activities associated with the Nationally Infectious Diseases list. As such, NNDSS is an umbrella term used to describe all public health infectious disease surveillance systems in this dissertation regardless of its specific type, usage, or source of origin.

(e.g., sales forecasting, product production, etc.), as well as those that described methods to determine the accuracy of forecasting models. Several statistical text books were studied that describe the use and application of Box-Jenkins.

To assure totality, a literature review was completed for the time period 1970–2012 utilizing the following databases: Scopus, PubMed, ProQuest, and Google Scholar. Searches utilized both individual and combinations of keywords using the terms: National Notifiable Disease Surveillance System, public health surveillance, public health data analysis, outbreak detection, predictive surveillance, public health informatics, Box-Jenkins forecasting, forecast modeling and accuracy, and forecast evaluation. An available feature within some of these databases allowed for the use of the “find similar article” feature that was used to identify additional articles. All citations were imported into an electronic database (RefWorks, Version 2.0).

Abstracts were reviewed for each identified article and full copies obtained if they contained information relevant for an in-depth evaluation. Key articles and their associated reference lists were also reviewed, and articles from these were reviewed and obtained as well. The majority of the literature review was completed during the first six months of 2012; a small portion was completed earlier during a pilot study that was used to determine the feasibility of the research questions associated with this dissertation. Eighty-one of the identified articles were used and/or cited in this dissertation.

Public Health Surveillance Systems

It is useful to understand the history of public health surveillance systems and how their data have been used have been used in order to appreciate how the forecasting principles presented in this dissertation may be used to identify potential outbreaks. In the book *Principles and Practice of Public Health Surveillance* (Teutsch & Churchill, 1994), Steven Thacker provides a historical review on the development of public health surveillance systems in the United States for the years 174 –1961. Highlights of this historical review include the first introduction of a surveillance system in Rhode Island in 1741, which required tavern keepers to report contagious disease among their patrons. In 1850, the federal government published the first mortality tables based upon death registration and decennial census data. Twenty-four years later, in 1874, the Massachusetts Board of Health established a surveillance system using a postcard based system to submit weekly disease reports from medical providers to the state health department. In 1878, Congress authorized the collection of morbidity data for use in quarantine control measures administered by the Public Health Service (PHS[2]). In 1893, Michigan became the first state to require medical providers to report infectious disease of public significance with the remainder of the states instituting similar systems within the next nine years. In 1914, there was a significant enhancement in disease reporting when the PHS

---

[2] The Public Health Service (PHS) was the federal agency responsible for enforcing quarantine measures in the United States. It was the forerunner of the Centers for Disease Control and Prevention (CDC).

assigned personnel to select state and local health departments to telegraph weekly disease reports to the PHS. This increased both the timeliness and completeness of infectious disease reporting from these jurisdictions. However, it was not until 1925 that all states participated in a nationally based reporting system (National Office of Vital Statistics, 1953). Expectations associated with public health surveillance at this time were limited to compiling and publishing morbidity statistics in weekly reports and were not used to actively identify outbreaks or other health related events.

In 1951, the Council of State and Territorial Epidemiologists (CSTE), in cooperation with the CDC, created a list of reportable diseases and established criteria that need to be in place for a disease to be considered reportable. For definitional purposes, the CSTE and the CDC define a notifiable disease as one for which regular, frequent, and timely information regarding individual cases is considered necessary for the prevention and control of the disease (Centers for Disease Control and Prevention, 2009). A case definition is uniform criteria for reporting cases (Centers for Disease Control and Prevention, 1997). An example of a case definition associated with Giardiasis is shown in Figure 1 (Council of State and Territorial Epidemiologists, 2012).

The list of national reportable diseases is contained in Appendix C, yet it should be noted that reporting is mandated only at the state or local level and is controlled by state legislation or local regulation (Centers for Disease Control and Prevention, 2012).

Case Definition: Giardiasis

Clinical Case Definition
An illness caused by the protozoan *Giardia lamblia* (aka G. *intestinalis* or G. *duodenalis*) and characterized by gastrointestinal symptoms such as diarrhea, abdominal cramps, bloating, weight loss, or malabsorption.

Laboratory criteria for diagnosis:
Laboratory-confirmed giardiasis shall be defined as the detection of *Giardia* organisms, antigen, or DNA in stool, intestinal fluid, tissue samples, biopsy specimens or other biological sample.

Case classification
Confirmed: a case that meets the clinical description and the criteria for laboratory confirmation as described above. When available, molecular characterization (e.g., assemblage designation) should be reported.

Probable: a case that meets the clinical description and that is epidemiologically linked to a confirmed case.

Figure 1: Example of a reportable disease case definition.

In the publication *Public Health Then and Now: Celebrating 50 Years of MMWR at CDC* (Centers for Disease Control and Prevention, 2011), Lisa Lee and Steven Thacker review the development of surveillance systems from 1961 through 2011. Highlights of this review include the deployment of a weekly telegraphic-based reporting system in 1961. This system essentially remained unchanged until 1975 when it was replaced with a telephone-based reporting system. In 1981, the telephone system began to support the electronic transfer of data associated with cumulative disease reports directly to computers at the CDC. The success of these data transfers led to the Electronic Surveillance Project (ESP) in 1984. The ESP was a five-year pilot project with the purpose of exploring issues associated with electronically transferring individual disease reports as opposed to the cumulative data being reported within the telephone-based system. The success of the ESP led directly to the development of the National Electronic Telephonic System for Surveillance (NETSS) in 1990. This system changed how reportable disease reports were sent to the CDC; prior to NETSS, data were reported as cumulative counts rather than individual case reports. Upon implementation of NETSS, states began electronically capturing and reporting individual case reports to CDC without personal identifiers (Centers for Disease Control and Prevention, 2009). The increase in the granularity of these reports (e.g., race, gender, exposure data, etc.) is now used by epidemiologists to help determine the source of an outbreak and for public health program evaluation purposes.

In the early 1990s additional development of NETSS allowed for the capture of expanded data sets associated with specific diseases (e.g., Lyme disease, vaccine preventable diseases, meningitis, etc.). By 1995, development of NETSS had reached the limit of its DOS-based architecture making the addition of other disease specific data sets impossible without completely rewriting the system. In response, a steering committee formed within CDC to investigate integrated public health surveillance systems. This committee produced a report, widely known as the “Katz report,” which served as the blueprint for one of the CDC’s new priority objectives, the creation of an integrated public health information and surveillance systems (Morris, Snider, & Katz, 1996).

In 2001, development begins on the National Electronic Disease Surveillance System (NEDSS). NEDSS is described by the CDC as an “internetbased infrastructure for public health surveillance data exchange." It is not a single application, but a system of interoperable subsystems and modules, based upon industry standards. It includes software applications developed by the CDC, state and local health departments, and those created by commercial vendors (Centers for Disease Control, 2008b). All NEDSS compliant systems developed since 2001 are built upon these agreed standards. These standards facilitate interoperability and simplify data transfer between disparate systems.

Nebraska began using the NEDSS Base System (NBS) in 2003 and within four years, 16 states had adopted the NBS as shown in Figure 2. From 2004 through the end of 2006, CDC received over 315,000 case reports from states using the NBS. The CSTE 2010 NEDSS Assessment Report (Council of State and Territorial Epidemiologists, 2010) showed that all states either had implemented the NBS or had developed or purchased a NEDSS compliant system. A breakdown of the states not using the NBS (shown as gray in Figure 2) show that 12 states had purchased a commercial off the shelf system, 15 states had develop a NEDSS compatible system in-house, 15 states used the CDC developed NBS, and eight states developed a hybrid system based upon customization of the NBS.

The use of NEDSS or NBS is a definite improvement for health departments and their associated disease surveillance activities. However, work on the analysis of data associated with these systems remains. For example, the CDC acknowledges that most outbreaks are identified in one of two ways with the first, and most common, being calls from a doctor, some other health care provider, or a citizen who knows of “several cases” (Centers for Disease Control and Prevention, n.d.). This method is often referred to as the astute observer. The second means of identifying outbreaks is the routine analysis of public health surveillance data. As the majority (i.e., 63%) of all the health departments in the United States serve jurisdictions less than 50,000 (Novich, 2011) and only 25% of all health departments employ an epidemiologist (Leep, 2007), a point forecast value and an associated forecast interval may be useful to evaluate the creditability of an astute observer report and the results of routine data evaluation.

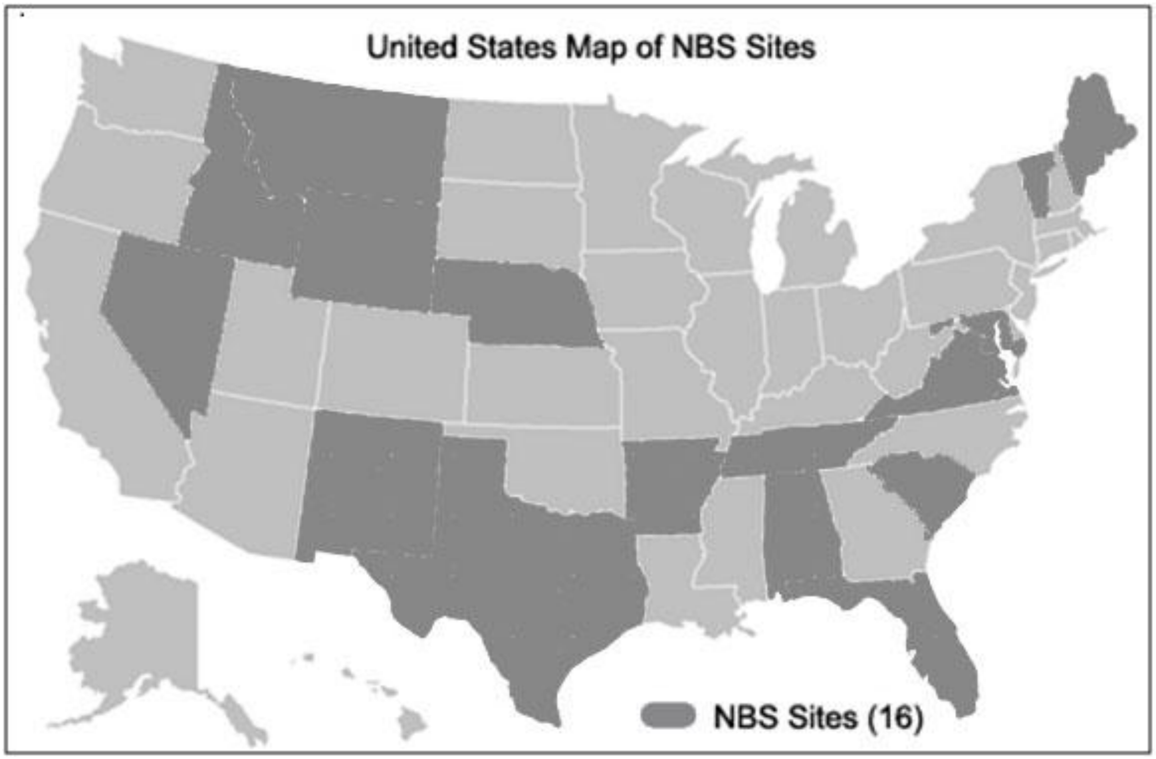


Figure 2: Location of states using the NEDSS Base System, 2012.

Analysis of National Notifiable Diseases Surveillance System Data

Disease control is a core function of public health and to accomplish this, public health practitioners routinely analyze public health data for a variety of purposes including the detection of unexpected increases in disease incidence that may indicate an outbreak or a change in disease patterns. The early detection of an outbreak may allow for the placement of effective interventions with the intent of mitigating excessive morbidity and mortality (Williamson & Weatherby, 1999). While the Box-Jenkins methodology has been used to forecast disease burden (Helfenstein, 1986) and within medical research (Helfenstein, 1996), its use with NEDSS data has been limited.

The analysis of data contained within NEDSS systems occur at the local, state, and federal level of public health practice, yet the literature for the most part is silent on these activities. As such, there are a limited number of peer-reviewed journal articles that describe efforts to apply regression and time-series analytical techniques to these data (Williamson & Weatherby, 1999). Most describe efforts aimed at developing systems to detect real time aberrations with Statistical Control Process (SPC) (Hutwagner, Maloney, Bean, Slutsker, & Martin, 1997) and their associated evaluation techniques based upon Shewhart Control Charts. Control charts are used to evaluate the stability of the process and variation represents a process that is considered out of control. An out of control finding would suggest the possibility of an outbreak. These studies are limited to a small subset of diseases. No peer-reviewed journal articles were identified describing the use of time-series analytical techniques for forecasting future disease burden based upon the use of NEDSS data. As such, the ongoing analysis of NEDSS data has a tendency to rely on straightforward temporal statistical methods such as historical monthly and weekly averages (Stroup, Wharton, Kafadar, & Dean, 1993).

Using historical data to produce any type of forecast carries an inherent level of uncertainty. To describe this uncertainty two types of information are needed; they are, a point forecast and a forecast interval (or confidence interval). Point forecasts are the best estimate of a future value and are easy to understand. However, by their nature point forecasts are incomplete since they describe only one possible outcome. The forecast interval is equally important as it describes the spread of the likely range, or potential distribution, of forecast outcomes (Cristofferson, 1998). Temporal statistical methods do not produce forecast intervals, thus making interpretation difficult beyond subjective comparisons between the observed and the expected data. Without a forecast interval, it is difficult to determine if an observed value is within an expected range of values or represents the potential beginning of an outbreak.

In the emerging field of syndromic surveillance, alternative methods of data analysis are being investigated for use within public health practice. While a specific definition for syndromic surveillance is lacking, these systems monitor surrogate data sources or disease related syndromes and not specific reportable diseases. Their intent is to monitor individual and/or population-based health indicators that may be detectable before confirmed laboratory diagnoses occur. The algorithms used within these systems are based upon symptoms or actions an ill person may exhibit prior at an actual diagnosis (Baer, Rodriguez, & Duchin, 2011). Examples of data used within these systems use include over-the-counter prescription sales, school absenteeism data, and syndrome categories including: fever, respiratory, gastrointestinal illness, hemorrhagic illness, localized cutaneous lesion, lymphadenitis, neurologic, rash, severe illness, and death (Henning, 2004). These systems are intended to support early outbreak detection by using near real-time reporting, automated outbreak identification, and related analytics (Chen, Zeng, & Yan, 2009), yet their usefulness, to date, remains unproven. While syndromic surveillance is potentially an important public health tool, the analysis of data associated with these systems is distinctly different from those associated with this dissertation as they do not rely upon diagnosed reportable disease data and their analytics based upon a cumulative sum chart (CUSUM), statistical control charts, and spatial analytical techniques (Kleinman, Abrams, Yih, Platt, & Kulldorff, 2006).

A detailed explanation of data analysis associated with historical averages is presented as it is a common method used to examine public health surveillance data. It is based upon the concepts associated with simple moving averages (Centers for Disease Control and Prevention, 2008a). As an example, a five-year monthly average, for a specific disease, for the month of October 2012 is the sum of the incident counts for the month of October for the years 2011, 2010, 2009, 2008, and 2007 and then divided by five. The resulting number represents a five-year monthly average. The use of a month as the unit of analysis is due to the relatively small incidence of disease within states that have small populations and within most local public health jurisdictions. Additional analysis takes the monthly expected averages and sums them by month to create an expected year-to-date (YTD) count. This allows the observed YTD count to be compared to the expected YTD count and a morbidity ratio calculated (Utah Department of Health, 2011). A morbidity ratio is simply the ratio of the observed counts divided by the expected counts of disease. Table 1 shows the results from this type of analyses.

Table 1*: Monthly Report of Notifiable Diseases, November 2011*

	Current Month # Cases	Current Month # Expected Cases (5-yr. Avg.)	# Cases YTD	# Expected YTD (5-yr Avg.)	YTD Morbidity Ratio (obs/exp)
Campylobacteriosis (*Campylobacter*)	15	20	413	321	1.3
Shiga toxin-producing *Escherichia coli* (E. coli)	4	7	166	113	1.5
Hepatitis A (infectious hepatitis)	0	1	6	10	0.6
Hepatitis B (serum hepatitis)	0	1	7	13	0.6
Meningococcal Disease	0	1	10	7	1.5
Pertussis (Whooping cough)	8	27	446	345	1.3
Salmonellosis (Salmonella)	16	24	293	304	1.0
Shigellosis (*Shigella*)	2	4	51	43	1.2
Varicella (Chickenpox)	13	72	318	630	0.5

Note. Utah Department of Health, Monthly Health Indicators Report, Nov. 2011. Source: http://health.utah.gov/opha/publications/hsu/1112_HlthSummit.pdf

The numerical data displayed in Table 1 represent point forecasts. As previously described, without forecast intervals, the interpretation of these data is subjected to interpretation by subjective judgment as there is no presentation of the distribution. The Morbidity Ratio is calculated in an attempt to counter this subjectivity, yet these results are potentially unstable especially when small numbers are involved.

At the CDC, and within larger public health jurisdictions, the analysis of NEDSS data occurs at a finer level of granularity utilizing individual weeks as the unit of analysis and uses methods in the calculation to help account for season variations in disease incidence over time (Centers for Disease Control and Prevention, 2006) The CDC presents the results of a five-year weekly average for publication in the *Morbidity and Mortality Weekly Report (MMWR)* Series by summing the incidence counts of the current month for the preceding five-year period; the sum is divided by five. A historical five-year weekly average is derived by summing the incidence counts of the current week, the two weeks prior to the current week, and the two weeks after the current week, for the preceding five-year period; the sum is divided by twenty-five.[3] As an example, a five-year weekly average, for a specific disease, for week number 38 of 2012 is the sum of the incidence counts for the weeks 36, 37, 38, 39, and 40 for the years 2011, 2010, 2009, 2008, and 2007 and then divided by 25. The resulting number

---

[3] These statistics are collected and compiled from reports sent by state and territorial health departments to the National Notifiable Diseases Surveillance System (NNDSS), which is operated by CDC in collaboration with the Council of State and Territorial Epidemiologists (CSTE).

represents a five-year weekly average. A visual representation of the five-year weekly average calculation method is shown in Table 2.

Table 2: *Five-Year Weekly Average Calculation*

Year	Week Number				
	Week 36	Week 37	Week 38	Week 29	Week 40
2011			Current Week		
2010	X1	X2	X3	X4	X5
2009	X6	X7	X8	X9	X10
2008	X11	X12	X13	X14	X15
2007	X16	X17	X18	X19	X20
2006	X21	X22	X23	X24	X25

Note. Five-year weekly average for current week = Sum of incidence counts X1 through X25, divided by twenty-five.

Literature to support the premise that NEDSS data are routinely analyzed by regression and time-series statistical techniques is scarce and is usually only alluded to in the literature. Upon reviewing original sources, it is apparent they describe only small-scale studies, focus on a limited number of diseases, and they are dated. For example, in the *History of Statistics in Public Health at CDC, 1960–2010: The Rise of Statistical Evidence* (Centers for Disease Control and Prevention, 2011), Donna Stroup and Rob Lyerla review the use of statistics at the CDC. In the introduction, they put forth the notion that “the use of statistics to assess data in epidemiology and public health are critical for identifying the causes of disease, modes of transmission, appropriate control and prevention measures, and for prioritizing and evaluating activities.” It is interesting to note that no mention is made of using statistics to identify outbreaks or to detect unusual patterns of disease. Later in the text, they do provide a limited discussion on statistics and surveillance in the following manner: “During this period [1980s], statistical methods for surveillance also advanced. The availability of methods for forecasting by using time series methods augmented previous regression results.” The first of the referenced materials associated with these statements describe a time-series analysis on two diseases and published in 1988 and the second on three diseases published in 1989 (Stroup, 1989).

Box-Jenkins Statistical Model

The Box-Jenkins approach to forecasting was first described by statisticians George Box and Gwilym Jenkins and was developed as a direct result of their experience with forecast problems in the business, economic, and control engineering applications (Box & Jenkins, 1994). The methods associated with Box-Jenkins resembles auto regression moving averages (ARMA) models with the exception that data within the time series has a steady underlying trend. Box-Jenkins accounts for this underlying trend by examining the differences between the successive observed values, instead of the values themselves. Box-Jenkins is one of several auto regressive integrated moving average (ARIMA) methods, all of which contain the following components (Caldwell, n.d.):

- Auto regression (AR): Regression that uses past values of itself to create forecasts instead of a predictor variable.

- Integrated (I): A time series has an underlying trend based upon differences between the successive observed values, instead of the values themselves. To retrieve the original data from the differences requires a form of integration.
- Moving Average (MA): The value in a time series forecast is influenced by the current error term and weighted error terms from the past.

The statistical theory behind Box-Jenkins is quite complicated and its use was somewhat limited until the relatively recent introduction of specialty forecasting packages that has greatly simplified its usage (Rycroft, 1995). Box-Jenkins forecasting is of greatest use when the underlying factors causing demand for products, services, revenue, and, in this case, disease burden are believed to behave in the future in much the same manner as it did in the past (Levenback & Cleary, 2006). A known shortcoming of Box-Jenkins forecasts is that they are based strictly upon univariate analysis, and this limits its use for exploring relationships to time and number of events (Pankratz, 1983). However, for this analysis the univariate methodology is appropriate as the intent is not to explore interdependent relationships but to explore the feasibility of using it to forecast disease burden.

Forecast Evaluation

While anyone can look retrospectively at a forecast and determine how accurate it was, it remains difficult to determine beforehand how accurate a specific forecast is going to be. There are a variety of statistical process tools that can be run when a given forecast is being prepared to help the forecaster determine how accurate a forecast may be. It is an accepted practice to test forecast accuracy beforehand by simulating trial forecast scenarios over a time period for which the actual results are known (Makridakis et al., 1993). When a forecaster has the time to conduct this type of analyses upon their forecasts prior to their actual usage, the use of a holdout sample is the preferred method as it provides a true test of the forecasts accuracy (Levenback & Cleary, 2006; Makridakis et al., 2000).

Holdout samples are simply a portion of the dataset that is withheld from the end of the series. These data are not used in the forecast model and are used to compare the actual value against a forecast value. Forecast accuracy may then be determined through a variety of methods (Hyndman, 2006) with the most common being the (a) Mean Absolute Deviation (MAD), (b) Mean Squared Error (MSE), and (c) Mean Absolute Percentage Error (MAPE). Each is discussed separately.

The MAD is a measure that is easy to calculate and the results are easily understood. It detects the average amount the forecast deviates from the actual data. It is most useful if the forecast will not be hurt by large errors. MAD is calculated by: $\mathrm{MAD} = \frac{{}_{t=1}^{n} et}{n}$ where et is the absolute value of the error term and *n* is the number of terms being evaluated.

The MSE is a more statistically based measure than MAD and is similar to evaluating the variance from random samples. MSE magnifies large errors found in forecasts and is useful when large forecasting errors are disruptive. MSE is calculated by: $\mathrm{MSE} = \frac{{}_{t=1}^{n} et2}{n}$. Both MAD and MSE are based upon the principle of making all errors positive either through the use of an absolute value or squaring the results of the errors. MAD has the advantage of being easier to explain where as MSE has the advantage of being easier to handle mathematically (Makridakis et al., 2000). Both MAD and MSE have a common limitation in that they each report errors in units independent of the actual data and provide no information about the magnitude of the errors.

MAPE is the most common measure of forecast accuracy and is used to measure the accuracy of forecasts associated with this dissertation. While a detailed explanation of MAPE is presented in the Methods chapter, a brief explanation of MAPE is provided here to contrast it with MAD and MSE.

MAPE calculates the mean of all the percentage errors for a given dataset without respect to the error being positive or negative in value and has the advantage of reporting the overall error as a percentage. For example, a MAPE result of .19 means the difference between the forecast value and the actual value is 19%, or put another way, the forecast is 79% accurate. MAPE is calculated as: $1 - \frac{MAPE}{1} \ \ where\ MAPE = \frac{1}{n} \quad {}_{t=1}^{n} \frac{At-Ft}{At}$. MAPE is sensitive to results with small or zero volume results as this skews the results.

Summary

The development of public health surveillance systems has created many independent systems. Within these systems are data that may be used to identify potential outbreaks, which is one of the stated reasons these systems exist.

Many continue to use relatively unsophisticated methods for the analysis of data contained within these systems. While there are many different methods that may be used to analyze these data, I have chosen to evaluate Box-Jenkins as the usefulness of it as an analytical tool for public health has not been systematically explored. Moreover, Box-Jenkins was selected as it supports a proactive approach to disease surveillance data analysis by allowing disease forecasting as opposed to simply looking at historical averages. MAPE is used to determine the overall accuracy of these forecasts as it is a recognized method to determine accuracy.

CHAPTER III

METHOD

In the preceding chapter, public health surveillance systems and the analysis of data associated with these systems were presented. Also discussed were the concepts of Box-Jenkins forecasting as well as providing a review of methods used to analyze these forecasts. This chapter presents the methods used to examine the study questions associated with this dissertation and includes: (a) the source of the data and an explanation of the data elements necessary to complete this analysis, (b) a description of the preparation necessary to prepare these data for the forecasting process, (c) a description of the mean absolute percentage error (MAPE), (d) the determination of forecast accuracy, and (e) the methods used to determine the forecast accuracy when they are stratified by jurisdictional size and rate of disease.

Data Source and Element Description

This dissertation utilizes NEDSS data obtained from the Utah Department of Health, Bureau of Epidemiology. While forecasts may be made with short time periods, longer time periods are preferred to as they will be more apt to capture trend, seasonality and other subtleties within the data (Stellwagen & Goodrich, 2008). As such, a 10-year dataset representing the years 2002–2011 is used. Although disease trends may change over time, the Box-Jenkins forecast model accounts for this by weighing values closer to the forecast event more heavily than those in the distant past. An additional factor in the selection of the 10-year dataset is that these data are contained within a single NEDSS based system.[4] Many data elements are contained within the available dataset, including demographics, disease specific risk factors, exposure history, and laboratory data, yet these data do not support forecasting, as forecasting is based upon an event of interest, time, and location. As such, to support this analysis, the following data elements are used: (a) disease type, (b) date of first report, and (c) jurisdiction of record. Each data element is discussed separately.

The disease type must be a disease contained on the list of Nationally Notifiable Diseases and Other Conditions of Public Health Importance**.** This list is maintained by the CDC and is updated on a yearly basis. Individual states modify this listing to support their unique needs; as such, diseases for this analysis data originate from a listing contained within the Utah Administrative Code, Communicable Disease Rule (Utah Administrative Code, 2012).

Each disease report has up to five dates that are relevant to the report or subsequent investigation. These dates are assigned to a hierarchy as follows: (a) date of disease onset, (b) date of diagnosis, (c) date of laboratory result, (d) date of first report to the public health system, and (e) MMWR report date. For this analysis, the date of first report to the public health system is used as all other dates are recorded either retrospectively or during the course of the

---

[4] The Utah Department of Health, Bureau of Epidemiology currently uses a NEDSS compliant system. This system replaced a NETSS based system in 2001. Data in the NETSS system are reported to be unreliable and/or incomplete.

disease investigative process making their accuracy suspect. Moreover, this date is the only one that is consistently recorded.

The jurisdiction of record represents the local public health jurisdiction from where the report originated or to the jurisdiction it was assigned to if the report was initially reported to the Utah Department of Health. It is the responsibility of the jurisdiction of record to investigate and record additional data associated with an investigation. These jurisdictions align with the geopolitical boundaries in Utah (i.e., counties). Six of the 12 local public health jurisdictions represent single counties with the remainder being multi-county jurisdictions; in addition, the entire State of Utah is considered a public health jurisdiction for this analysis. All of the public health jurisdictions are shown in Figure 3. The populations within these jurisdictions range from 23,530 to 2,763,885.

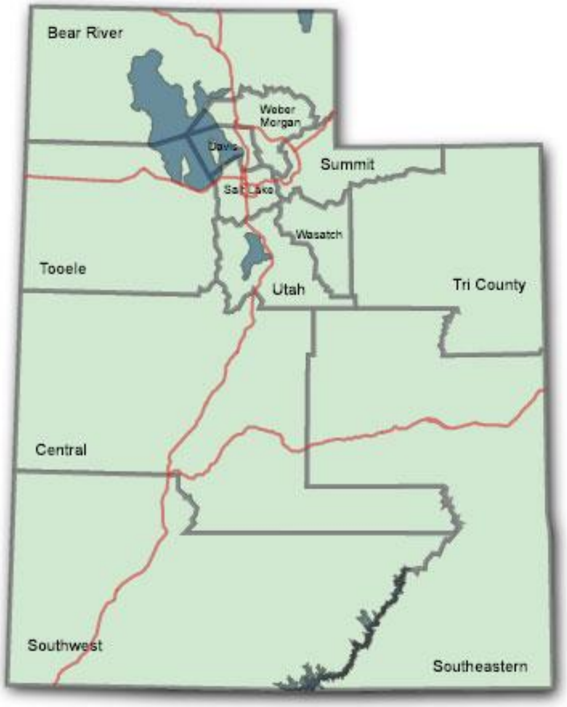


Figure 3: Utah Public Health Department jurisdictions.

Data Preparation

To prepare these data for forecasting, the entire dataset is stratified by year (*n* = 10), disease type (*n* = 50[5]) and public health jurisdiction (*n* = 13[6]). Rates for each disease are also calculated utilizing a mid-point population estimate to control for the 23.8% population increase over the time period of the study (U.S. Census Bureau, 2012*)*. For a disease to be included in this analysis, the following criteria must be met:

- A disease must have a 10-year average rate greater than 0.16/100,000. This rate equates to approximately five cases a year for the entire state.
- A disease must be listed within the Communicable Disease Rule for the entire 10-year period associated with this study. Several diseases were either added or removed from the listing over the 10-year period associated with this study; as such, they are excluded from this analysis.
- A disease that has a mandatory telephone reporting requirement is excluded. This exclusion is based upon potential inconsistencies with the date of first report. Moreover, these diseases typically represent rare events and often have a limited number of reports (i.e., rate < 0.16/100,000).

---

[5] Although there are 75 diseases listed within the Utah Communicable Disease Rule, many did not have a single confirmed laboratory report in the 10-year dataset used in this study.

[6] There are 12 local public health jurisdictions in Utah; for this study, the 13th encompasses all of the local public health jurisdictions.

Diseases selected for inclusion are shown in Figure 4.

Disease Name

Amebiasis
Campylobacter
Chlamydia trachomatis
Coccidioidomycosis
Dengue
E. coli
Encephalitis
Giardiasis
Gonorrhea
Haemophilus influenzae
Hepatitis B (acute)
Hepatitis C (acute)
Legionellosis
Lyme disease
Malaria
Meningitis (viral)
Pertussis
Rocky Mountain spotted fever
Salmonellosis (excluding Typhoid)
Shigella
Streptococcus pneumoniae
Varicella

Figure 4. Infectious diseases selected for inclusion in determination of accuracy of Box-Jenkins forecasting methods.

Upon completion of the disease inclusion assessment, preliminary disease specific forecasts are conducted. All forecasts use fitted values that are based upon the weighted average of the previous observations, plus a combination of weighted error values associated with the observations (Box, Jenkins, & Bacon, 1967). This results in events that are closer to the forecast being weighted more heavily than events in the distant past.

The fitted value, used to generate the preliminary forecast, is screened for outliers. Forecasting using fitted values that contain outliers has the potential to substantially influence the accuracy of a forecast. A solution is to screen the historical data for outliers and replace them with adjusted values prior to generating the actual forecasts (Stellwagen & Goodrich, 2008). In this case, an outlier is defined as a value in the historical data that is greater than three standard deviations from the fitted value used to produce the forecast models. Data that meet this definition are adjusted to a corrected value that is associated with the upper limit of one standard derivation. This allows the impact of the outlier to be accounted for but reduces the overall influence on the forecast. This outlier identification and correction process is presented in graphic detail in Appendix D**.** All forecasting is completed using Forecast PRO XE (Forecast Pro, Version 5.5). The modified dataset is now ready for use to answer the research question associated with this dissertation:

1. Can Box-Jenkins forecasts produce disease specific forecasts that are equal to or greater than 75% accurate?
2. For diseases specific forecasts that are equal to or greater than 75% accurate, what influence does jurisdictional size and rate of disease have upon the accuracy of the forecasts?

Mean Absolute Percentage Error

The accuracy of a forecast is the quantified difference between the forecast value and the actual value for a defined time period. In this case, observed values are the incidence of disease. By convention, forecast values are subtracted from the actual value as shown in the following formula: $Et = Yt - Ft$ where *E* is the forecast error at period *t*, *Y* is the actual value for period *t,* and *F* is the forecast value for period *t*.

Additional calculations are necessary to determine the overall accuracy of a forecast using the mean absolute percentage error (MAPE) calculation. Forecast accuracy is determined by comparing the forecast values against the actual values within the holdout samples by calculating a MAPE for each of the forecasts trials. MAPE is defined as the average of percentage errors and is calculated as: $1 - \frac{MAPE}{1} \;\; where\, MAPE = \frac{1}{n} \;\; {}^{n}_{t=1} \frac{At-Ft}{At}$. By convention *A* is the actual value, *F* is the forecast value and *t* is a period in time. Once MAPE is determined, it is then possible to derive the absolute accuracy of the forecast using the following calculation $1 - MAPE$.

A limitation of MAPE is when the forecast model perfectly predicts the actual value as this result is a situation requiring division by zero. To account for this possibility, perfect predictions will be removed from the overall accuracy determination and reported separately.

Determination of Forecast Accuracy

As a retrospective secondary data analysis, the results from the Box-Jenkins forecasts are compared to actual data from the same time period. To accomplish this, disease specific forecasts are made against three holdout samples with each being one calendar year in duration. A holdout sample is a portion of the dataset reserved for evaluating the accuracy of a forecast. The underlying logic of using holdout samples to determine accuracy is that portions of the time series data are withheld before the forecasts are created. This allows forecast results to be compared against data which have been withheld rather than trying to determine how well they perform against data which has been used to create the forecast. This allows for a simple comparison on their forecasting accuracy by comparing their MAPE performance against their associated holdout sets. Each disease selected for inclusion will have three forecasts completed and the results averaged. This will be completed by using data from years 2002–2008 to forecast year 2009 and the accuracy determined. The process is repetitive with data from 2002–2009 used to forecast year 2010, and data from 2002–2010 used to forecast year 2011.

There are 22 diseases eligible for forecasting based upon the established inclusion criteria as outlined previously. Each disease will have three forecast trials with the results averaged. This will result in 66 forecasts trials and 22 averaged results. The threshold for forecasting accuracy is set at 75% with forecasts meeting or exceeding this threshold considered accurate and forecasts below this threshold considered inaccurate. The 75% threshold was determined by consulting with state and local public health officials. Overall accuracy of the Box-Jenkins ability to forecast future disease burden is determined by comparing the number of accurate forecasts against the number of inaccurate forecasts.

Forecast Accuracy Based Upon Jurisdiction Size and Rate of Disease

All forecasting models have a point upon which the lack of historical data affects its ability to create accurate forecasts (Boylan, 2005). Public health jurisdictions vary in size and the populations and in this study they have over a one hundred-fold difference ranging from 23,530 to 2,763,885. As a function of size, public health jurisdictions with smaller populations have a fewer number of reportable diseases. An additional factor influencing the number of reports is that the incidence of reports varies by disease. These two factors often results in smaller jurisdictions having no reports for specific diseases for extended periods of time. It is therefore necessary to examine forecast results by jurisdictional size and rate of disease to determine at which point the Box-Jenkins forecast model fails.

Disease specific forecasts that are equal to or greater than 75% accurate will be evaluated by examining the relationship between each holdout sample's actual value and the forecast interval associated with each disease specific forecast. The forecast intervals are calculated at one, two, and three standard deviations. To quantify the results, the following rule set is used:

- The forecast is acceptable if the actual monthly value does not exceed the upper limit of the forecast's predication interval more than three times in the 36-month holdout period.[7]

---

[7] This rule will be adjusted if a forecast cannot be completed for a specific year. For example, if only two forecast years are completed for a specific disease due to sparse data then the rule would be applied as follows: The forecast is acceptable if the actual monthly value does

- Actual monthly values greater than those associated with the third forecast interval are classified as outbreaks.
- If there are less than three actual monthly values during a given year, a forecast will not be completed for that specific disease within that jurisdiction and will be reported as incomplete.

To examine the effect of jurisdictional size and rate of disease have upon forecast accuracy, it is necessary to select jurisdictions with differing population sizes. Seven public health jurisdictions are used to represent these differing population sizes. Selection for inclusion is based upon the following methodology. The first category is the largest public health jurisdiction. The population of this jurisdiction is then divided in half and the jurisdiction with the population closest to the product is selected for inclusion. This process is repetitive until the final jurisdiction is selected. Table 3 shows the public health jurisdictions by population selected for inclusion.

Statistical Testing

Beyond the determination of the absolute accuracy of the forecasts using MAPE statistical testing is used to examine the various relationships between jurisdictional size, rate of disease, forecast accuracy, and forecast intervals. These tests will be used to help determine usefulness of Box-Jenkins as a tool

---

not exceed the upper limit of the forecast’s predication interval more than two times in the 24-month holdout period.

for public health and help determine the point where Box-Jenkins forecasts begin to fail. All statistical tests are conducted using SPSS (SPSS, Version 21.0).

Table 3: *Public Health Jurisdiction by Population*

Public Health Jurisdiction	Counties	Population	Designation	Inclusion
State of Utah	All Counties	2,763,885	Large	Yes
Salt Lake	Salt Lake	1,029,655	Large	Yes
Utah	Utah	516,564	Large	Yes
Davis	Davis	306,479	Medium	No
Weber-Morgan	Weber, Morgan	240,705	Medium	Yes
Southwest	Beaver, Garfield, Iron, Kane, Washington	203,204	Medium	No
Bear River	Box Elder, Cache, Rich	167,895	Medium	Yes
Central Utah	Juab, Millard, Piute, Sanpete, Sevier, Wayne	75,707	Medium	Yes
Tooele	Tooele	58,218	Medium	No
Southeastern	Carbon, Emery, Grand, San Juan	56,350	Medium	No
Summit	Summit	36,324	Small	Yes
Wasatch	Wasatch	23,530	Small	No

Note. The Large, Medium, Small designation matches the criteria used in the 2010 National Profile of Local Health Departments.

Pearson correlation is used to examine the following relationships as these data have a normal distribution and are continuous:

- The relationship between rate of disease and forecast accuracy (both grouped [all disease forecasts] and for diseases with > 75% accuracy).
- The relationship between jurisdictional size[8] and forecast accuracy.
- The relationship between incomplete forecasts and jurisdictional size.
- The relationship between incomplete forecasts rate of disease.
- The relationship between rate of disease and number of observations in the first forecast interval.

---

[8] Jurisdictional size is a continuous variable as it is based upon the population of the jurisdiction.

CHAPTER IV

RESULTS

This chapter is presented in two sections; each contains the results associated with the research questions in this dissertation. The first section reports on the global accuracy of the Box-Jenkins forecasting trials. The second section reports on the influence jurisdictional size and rate of disease has upon forecast accuracy and forecast intervals. Figure 5 shows the workflow associated with the analysis and subsequent reporting of these data.

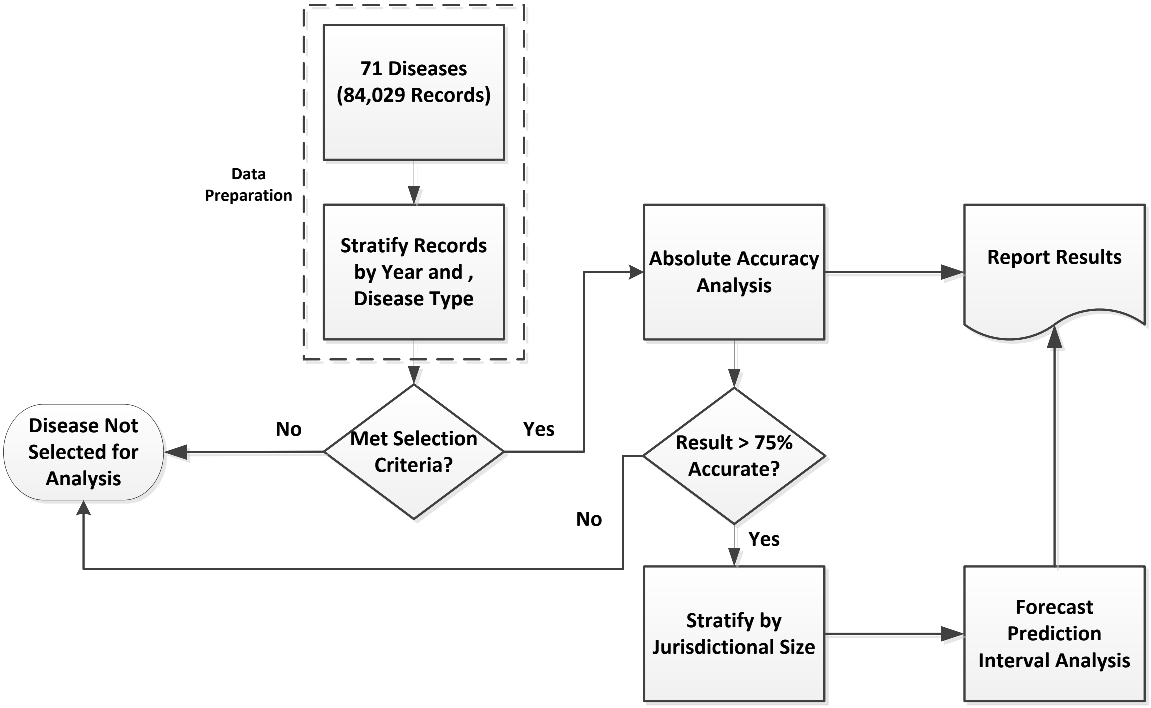


Figure 5: Disease forecast analysis and reporting workflow.

Twenty-two diseases were initially selected for analysis using methods outlined in the previous chapter. During the analysis, six were excluded: three for data inconsistencies; two for incomplete datasets; and one for a change in the laboratory component of the case definition that resulted in a large increase in the incidence of reported disease. As a result, 16 diseases are reported on in this chapter.

Forecast Accuracy

Utilizing data representing all public health jurisdictions in Utah, three forecast trials were completed for each disease resulting in 48 trials. Within these trials there are 576 monthly forecasts. There are 53 perfect predictions that were removed from the absolute accuracy calculations. This is necessary as a perfect forecast results in a situation that would require division by zero in the MAPE calculation. The forecast results charts are contained in Appendix E.

The forecasts trial results were grouped by disease, compared against their associated holdout sample,[9] and absolute accuracy calculated. In aggregate, the absolute accuracy of all forecast trails[10] is 71% (range: 43.4 – 91.7%). Ten of the 16 disease forecasts (63%) had an absolute accuracy greater than 75%, four (25%) had an absolute accuracy between 52.6% and 69.6%, and

---

[9] Holdout samples are simply a portion of the dataset that is withheld from the end of the series. These data are not used in the forecast model and are used to compare actual value against the forecast value.

[10] The aggregate absolute accuracy results do not contain the forecasts that were perfect (*n* = 53).

two (12%) had an absolute accuracy of less than 50%. Table 4 shows the absolute accuracy results for all public health jurisdictions in Utah.

Table 4: *Forecast Trial Results: Absolute Accuracy of Disease for All Public Health Jurisdictions in Utah*

Disease Name	Disease Rate	Forecast Accuracy
Amebiasis	0.4/100K	85.2%
Campylobacter	13.6/100K	78.9%
*Chlamydia trachomatis****	183/100K	n/a
Coccidioidomycosis	2.3/100K	75.4%
Dengue	0.2/100K	79.8%
E. coli*	4.7/100K	n/a
Encephalitis	0.4/100K	91.7%
Giardiasis	15.2/100K	75.3%
Gonorrhea*	25.4/100K	n/a
Haemophilus influenza	1.3/100K	58.8%
Hepatitis B (acute)*	1.7/100K	n/a
Hepatitis C (acute)**	17.7/100K	n/a
Legionellosis	1.0/100K	63.8%
Lyme disease	0.9/100K	81.8%
Malaria	0.4/100K	78.3%
Meningitis (viral)	4.0/100K	62.5%
Pertussis	17.1/100K	60.0%
Rocky Mountain spotted fever**	0.4/100K	n/a
Salmonellosis	12.6/100K	76.8%
Shigella	2.0/100K	43.4%
Streptococcus pneumonia	5.7/100K	76.5%
Varicella	23.0/100K	47.2%

* Analysis not completed due to inconsistencies within dataset
** Analysis not completed due to incomplete dataset
*** Analysis not completed due to change in case definition / laboratory test

Influence of Jurisdictional Size and Rate of Disease on Forecast Accuracy

The 10 diseases with forecast accuracy greater than 75% were examined to determine the possible influence jurisdictional size and rate of disease has upon a disease's actual location within the first, second, or third forecast interval. While technically 100% accurate, forecasts that predict zero events and contained a forecast interval of zero were removed from this portion of the analysis. This was done to unmask the influence these forecasts have upon accuracy and to help determine the point when the forecasting model fails. In some cases, the majority of forecast results for a specific disease were removed. As such, the results associated with forecast accuracy in this portion of the analysis are considerably lower than the aggregate results described above.

Three forecast trials were completed for each disease (*n* = 10) and jurisdiction (*n* = 7) resulting in 210 additional forecasts. Within these trails there are 2, 520 monthly forecasts. These forecasts produced usable results in 53.6% (*n* = 1,350) of the trails at one, two and three forecast intervals as shown in Table 5.

The relationship between jurisdictional size and forecast accuracy was examined using Pearson correlation. There is a moderate, correlation between the two variables (*r* = –.396, *n* = 7, *p* = .380). The association is not statistically significant; therefore, jurisdictional size is not related to the accuracy of a forecast.

The relationship between rate of disease and forecast accuracy was investigated using Pearson correlation. There is a moderate, negative correlation between the two variables (*r* = –.496, *n* = 10, *p* = .164]. The association is not statistically significant; therefore, rate of disease is not related to the accuracy of a forecast.

Table 5: *Forecast Accuracy and Actual Values Location in the Forecast Intervals*

Jurisdiction (population)	Disease Rate per 100K	Accuracy	Incomplete Forecast	Outbreak	1st Forecast Interval	2nd Forecast Interval	3rd Forecast Interval
Giardiasis							
State of Utah (2,763,885)	15.2	75.3%	0	0	72.2%	97.2%	100%
Salt Lake Valley (1,029,655)		53.3%	0	0	80.5%	100%	100%
Utah County (516,564)		32.5%	0	0	72.2%	97.2%	100%
Weber/Morgan (240,705)		45.9%	0	0	94.4%	100%	100%
Bear River (167,895)		85.1%	0	0	94.4%	100%	100%
Central Utah (75,707)		52.5%	0	0	94.4%	100%	100%
Summit (36,324)		82.5%	0	2	94.4%	94.4%	94.4%
				Mean:	84.1%	98.4%	99.2%
Campylobacter							
State of Utah (2,763,885)	13.6	78.9%	0	2	63.9%	88.9%	88.9%
Salt Lake Valley (1,029,655)		69.6%	0	3	72.2%	88.9%	91.7%
Utah County (516,564)		54.8%	0	4	52.8%	77.8%	88.9%
Weber/Morgan (240,705)		52.6%	0	0	77.8%	94.4%	100%
Bear River (167,895)		52.2%	0	1	80.5%	97.2%	97.2%
Central Utah (75,707)		63.4%	0	1	97.2%	97.2%	97.2%
Summit (36,324)		84.5%	0	0	91.6%	97.2%	100%
				Mean:	84.1%	98.4%	99.2%

Table 5—Continued

Jurisdiction (population)	Disease Rate per 100K	Accuracy	Incomplete Forecast	Outbreak	1st Forecast Interval	2nd Forecast Interval	3rd Forecast Interval
Salmonella							
State of Utah (2,763,885)	12.6	76.8%	0	0	76.5%	94.4%	100%
Salt Lake Valley (1,029,655)		40.6%	0	0	63.8%	91.7%	100%
Utah County (516,564)		33.4%	0	1	58.3%	91.7%	97.2%
Weber/Morgan (240,705)		45.9%	0	1	83.3%	94.4%	94.4%
Bear River (167,895)		51.6%	0	0	69.4%	100%	100%
Central Utah (75,707)		83.7%	0	1	94.4%	97.2%	97.2%
Summit (36,324)		64.1%	0	0	94.4%	100%	100%
				Mean:	76.6%	91.7%	94.8%
Streptococcus pneumonia							
State of Utah (2,763,885)	5.7	76.5%	0	0	66.7%	94.4%	100%
Salt Lake Valley (1,029,655)		58.4%	0	0	61.1%	94.4%	100%
Utah County (516,564)		57.3%	0	1	83.3%	100%	100%
Weber/Morgan (240,705)		60.2%	0	1	66.7%	88.9%	94.4%
Bear River (167,895)		62.9%	0	2	61.1%	94.4%	94.4%
Central Utah (75,707)		n/a	36	0	n/a	n/a	n/a
Summit (36,324)		n/a	36	0	n/a	n/a	n/a
				Mean:	67.8%	94.4%	97.8%

Table 5—Continued

Jurisdiction (population)	Disease Rate per 100K	Accuracy	Incomplete Forecast	Outbreak	1st Forecast Interval	2nd Forecast Interval	3rd Forecast Interval
Coccidioidomycosis							
State of Utah (2,763,885)	2.3	75.4%	0	0	94.4%	100%	100%
Salt Lake Valley (1,029,655)		84.4%	0	0	94.4%	100%	100%
Utah County (516,564)		n/a	36	0	n/a	n/a	n/a
Weber/Morgan (240,705)		n/a	36	0	n/a	n/a	n/a
Bear River (167,895)		n/a	36	0	n/a	n/a	n/a
Central Utah (75,707)		n/a	36	0	n/a	n/a	n/a
Summit (36,324)		n/a	36	0	n/a	n/a	n/a
				Mean:	94.4%	100%	100%
Lyme							
State of Utah (2,763,885)	0.9	81.8%	0	0	83.3%	94.4%	100%
Salt Lake Valley (1,029,655)		50.0%	24	0	50.0%	100%	100%
Utah County (516,564)		n/a	36	0	n/a	n/a	n/a
Weber/Morgan (240,705)		n/a	36	0	n/a	n/a	n/a
Bear River (167,895)		n/a	36	0	n/a	n/a	n/a
Central Utah (75,707)		n/a	36	0	n/a	n/a	n/a
Summit (36,324)		n/a	36	0	n/a	n/a	n/a
				Mean:	66.7%	97.2%	100%

Table 5—Continued

Jurisdiction (population)	Disease Rate per 100K	Accuracy	Incomplete Forecast	Outbreak	1st Forecast Interval	2nd Forecast Interval	3rd Forecast Interval
Encephalitis							
State of Utah (2,763,885)	0.4	97.7%	0	0	97.2%	97.2%	100%
Salt Lake Valley (1,029,655)		65.6%	0	0	91.7%	91.7%	100%
Utah County (516,564)		16.7%	24	0	91.7%	91.7%	100%
Weber/Morgan (240,705)		91.7%	12	4	83.3%	83.3%	83.3%
Bear River (167,895)		n/a	36	0	n/a	n/a	n/a
Central Utah (75,707)		n/a	36	0	n/a	n/a	n/a
Summit (36,324)		n/a	36	0	n/a	n/a	n/a
				Mean:	91.0%	91.0%	95.8%
Malaria							
State of Utah (2,763,885)	0.4	78.3%	0	0	94.4%	100%	100%
Salt Lake Valley (1,029,655)		77.8%	0	1	97.2%	97.2%	97.2%
Utah County (516,564)		n/a	36	0	n/a	n/a	n/a
Weber/Morgan (240,705)		n/a	36	0	n/a	n/a	n/a
Bear River (167,895)		n/a	36	0	n/a	n/a	n/a
Central Utah (75,707)		n/a	36	0	n/a	n/a	n/a
Summit (36,324)		n/a	36	0	n/a	n/a	n/a
				Mean:	95.8%	98.6%	98.6%

Table 5—Continued

Jurisdiction (population)	Disease Rate per 100K	Accuracy	Incomplete Forecast	Outbreak	1st Forecast Interval	2nd Forecast Interval	3rd Forecast Interval
Amebiasis							
State of Utah (2,763,885)	0.4	85.2%	0	0	91.7%	97.2%	100%
Salt Lake Valley (1,029,655)		62.8%	0	0	94.4%	100%	100%
Utah County (516,564)		n/a	36	0	n/a	n/a	n/a
Weber/Morgan (240,705)		n/a	36	0	n/a	n/a	n/a
Bear River (167,895)		n/a	36	0	n/a	n/a	n/a
Central Utah (75,707)		n/a	36	0	n/a	n/a	n/a
Summit (36,324)		n/a	36	0	n/a	n/a	n/a
				Mean:	93.1%	98.6%	100%
Dengue							
State of Utah (2,763,885)	0.2	79.8%	0	0	94.4%	100%	100%
Salt Lake Valley (1,029,655)		n/a	36	0	n/a	n/a	n/a
Utah County (516,564)		n/a	36	0	n/a	n/a	n/a
Weber/Morgan (240,705)		n/a	36	0	n/a	n/a	n/a
Bear River (167,895)		n/a	36	0	n/a	n/a	n/a
Central Utah (75,707)		n/a	36	0	n/a	n/a	n/a
Summit (36,324)		n/a	36	0	n/a	n/a	n/a
				Mean:	94.4%	100%	100%
				Totals:	84.1%	96.5%	98.5%

Forecast Acceptance

Using the rule set defined in the methods section, 84.1% (*n* = 1,135) of the monthly forecasts were determined to be acceptable at the first forecast interval, 87.5% (*n* = 1,181) were acceptable at the second forecast interval, and 98.7% (*n* = 1,332) acceptable at the third forecast interval. The remaining (*n* = 21) actual values observed in the holdout samples were classified as outbreaks (i.e., > third forecast interval). It is interesting to note that when an actual value is located within the second forecast interval, it progresses to outbreak status 54% (*n* = 21) of the time.

The number of acceptable forecasts located in the first forecast interval is 16% greater than expected when considering results in a normal distribution or bell curve. The excess numbers of forecasts contained in the first forecast interval are from values one would expect to find in the second forecast interval. The relationship between the percentage of forecast values contained within the first forecast interval and rate of disease was investigated using Pearson correlation. There is a moderate, negative correlation between the two variables ($r = –.328$, $n = 10$, $p > .05$). The association is not statistically significant; therefore, the rate of disease is not associated with the number of forecast values located within the first forecast interval.

Model Failure Point

The failure point of the model is due to the lack of available data and not upon the models ability to produce forecasts. Even forecasts that are not overly accurate have usable results that generally fall within the first or second forecast interval. Table 6 shows the point where data used in this study failed to produce results. It is based upon the rate of disease and jurisdictional size.

Table 6: *Forecast Model Failure Point by Rate of Disease and Jurisdictional Size*

	State of Utah (2,763,885)	Salt Lake Valley (1,029,655)	Utah County (516,564)	Weber/ Morgan (240,705)	Bear River (167,895)	Central Utah (75,707)	Summit (36,324)
Giardiasis 15.2/100K	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Campylobacter 13.6/100K	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Salmonella 12.8/100K	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Strep pneumoniae 5.7/100K	Pass	Pass	Pass	Pass	Fail	Fail	Fail
Coccidioidomycosis 2.3/100K	Pass	Pass	Fail	Fail	Fail	Fail	Fail
Lyme 0.9/100K	Pass	Pass	Pass	Fail	Fail	Fail	Fail
Encephalitis 0.4/100K*	Pass	Fail	Fail	Fail	Fail	Fail	Fail
Malaria 0.2/100K	Pass	Fail	Fail	Fail	Fail	Fail	Fail

*There are three diseases at the 0.4/100K level; each has the same failure point.

The relationship between the number of incomplete forecasts and jurisdictional size was investigated using Pearson correlation. There is a strong, negative correlation between the two variables (r = –.936, n = 10, p = .01), with smaller jurisdictions having more incomplete forecasts than larger jurisdictions. In support of this finding, there is a strong negative correlation between the rate of disease and incomplete forecasts (r = –.961, n = 10, p = .01) with diseases associated with lower rates having more incomplete forecasts than diseases with larger rates.

Perfect Forecasts

There were 53 perfect forecasts and they were removed from the MAPE calculation as perfect forecasts require division by zero in a MAPE calculation, thus nullifying the use of MAPE in the determination of the overall accuracy of the forecasts. However, these forecasts do need to be accounted for in this analysis. Table 7 shows the distribution of these perfect forecasts.

Table 7: *Number of Perfect Forecasts by Rate of Disease and Jurisdictional Size*

Disease Name	Disease Rate	Forecast Accuracy	Number of Perfect Forecast
Amebiasis	0.4/100K	85.2%	6
Campylobacter	13.6/100K	78.9%	1
Coccidioidomycosis	2.3/100K	75.4%	3
Dengue	0.2/100K	79.8%	5
Encephalitis	0.4/100K	91.7%	5
Giardiasis	15.2/100K	75.3%	2
Haemophilus influenza	1.3/100K	58.8%	3
Legionellosis	1.0/100K	63.8%	4
Lyme disease	0.9/100K	81.8%	4
Malaria	0.4/100K	78.3%	6
Meningitis (viral)	4.0/100K	62.5%	5
Pertussis	17.1/100K	60.0%	1
Salmonellosis	12.6/100K	76.8%	0
Shigella	2.0/100K	43.4%	4
Streptococcus pneumonia	5.7/100K	76.5%	3
Varicella	23.0/100K	47.2%	1

The relationship between rate of disease and the number of perfect forecasts was investigated using Pearson’s correlation. There is a strong, negative correlation between the two variables (r = –.588, n = 16, p = .05), with a low rate of disease associated with an increase number of perfect forecasts. The association is statistically significant.

The relationship between forecast accuracy and the number of perfect forecasts was investigated using Pearson's correlation. There is a moderate correlation between the two variables (r = .348, n = 16, p = .186). The association is not statistically significant; therefore, forecast accuracy is not related to the number of perfect forecasts.

Summary

The cumulative accuracy of all the forecast trials was 71% with a range of 43.4 – 91.7%. Although lower than the 75% threshold associated with the original research question, it was found that 63% of the disease specific forecasts had accuracies greater than 75%. Thus, it is valid to state that Box-Jenkins forecasts can produce disease specific forecasts that are equal to or greater than 75% accurate but not for all diseases of public health significance. There was no statistical relationship between rate of disease and forecast accuracy when examining the results in aggregate and when using only the disease forecasts that were greater than 75% accurate.

There was no statistical relationship noted between jurisdictional size and accuracy as well as disease rate and accuracy; therefore, it is valid to state that jurisdictional size and rate of disease are not associated with the accuracy of Box-Jenkins disease specific forecasts.

A comparison of the forecast intervals against the holdout values found a disproportionate number (84%) of actual values located in the first forecast interval. The number of acceptable forecasts located in the first forecast interval is 16% greater than expected when considering results in a normal distribution or bell curve. The excess number of forecasts contained in the first forecast interval is from values one would expect to find in the second forecast interval. There was no statistical relationship between the number of forecast values contained within the first forecast interval and rate of disease. It was also found that when an actual is located within the second forecast interval, it progresses to outbreak status 54% of the time.

The failure rate of the Box-Jenkins forecasts is based upon the rate of disease and jurisdictional size and not upon the ability of the model to create accurate forecasts. Both low rate of disease and small jurisdictional size were significantly more likely to be unable to produce forecasts due the number of incomplete reports in the historical data.

Lastly, there was statistical relationship noted between disease rate and the number of perfect forecasts, but this did not hold true for the relationship between forecast accuracy and the number of perfect forecasts.

CHAPTER V

DISCUSSION

The threat from new and emerging diseases as well as those associated with bioterrorism has focused attention on public health surveillance systems. Health departments across the nation are enhancing existing surveillance systems or developing new ones to better detect disease trends and potentially identify outbreaks earlier. However, methods that support the analysis of data within these systems, in many cases, has not progressed beyond rudimentary methods such as historical counts of disease or comparisons based upon simple moving averages. Regardless of the method, detection algorithm, or statistical model, all outbreak detection methodologies are based upon the premise of comparing the observed count against the expected count of disease. The creation of the expected count, or baseline, is challenging as many infectious disease patterns are not stable due to increasing or decreasing populations, seasonality of disease, the occurrence of outbreaks and other external conditions.

This study specifically examined the use of Box-Jenkins forecasting and tested its ability to create accurate baselines. Specifically, as a retrospective secondary data analysis using NEDSS data from a state-based disease surveillance system this study examined the use of Box-Jenkins forecasts models and its ability to create accurate baselines for use in public health practice. To accomplish this task this study posed two research questions, they are:

1. Can Box-Jenkins forecasts produce disease specific forecasts that are equal to or greater than 75% accurate?
2. For disease specific forecasts that are equal to or greater than 75% accurate, what influence does jurisdictional size and rate of disease have upon the accuracy of the forecasts?

Forecast Accuracy

Forecast accuracy was determined by calculating the mean absolute percentage error or MAPE associated with each disease specific forecast. MAPE is the most common measure of forecast accuracy and determines error by calculating the mean of all the percentage errors for a given dataset. It accomplishes this without respect to the error being positive or negative in value and has the advantage of reporting the overall error as a percentage. Initially, 22 diseases were selected for review representing a range of disease rates. Six were unable to be evaluated due to a variety of issues, which underscores the challenges of working with secondary data sets.

Utilizing data representing all public health jurisdictions in Utah, three forecast trials were completed for each disease resulting in 48 trials. Within these trials there were 576 monthly forecasts. By calculating MAPE for each of these forecasts, it was determined that the cumulative accuracy of all the forecast trials was 71% with a range of 43.4 – 91.7%. Although the cumulative accuracy was lower than the 75% threshold associated with the original research question, it was found that 63% of the disease specific forecasts had accuracies greater than 75%. Thus, it is valid to state that Box-Jenkins forecasts can produce disease specific forecasts that are equal to or greater than 75% accurate but not for all diseases of public health significance.

While some may argue that using a 75% threshold for accuracy determination is arbitrary, it was necessary to set a value that would allow comparisons to be made. Additionally, the 75% threshold was established through a consultative effort with individuals actively engaged in public health practice.

There was no statistical relationship between rate of disease and forecast accuracy when examining the results in aggregate and when using only the disease forecasts that were greater than 75% accurate. This implies that the rate of disease has no effect on Box-Jenkins forecast accuracy. However, caution should be used with the interpretation of this finding; while insignificant, both statistical tests reported moderate negative correlations suggesting that diseases with small rates were more accurate than diseases with large rates. These results may be simply a function of the number of diseases examined in this study.

Forecast results are estimates of future values and are easy to understand. However, by their nature, forecast values are incomplete since they describe only one possible outcome. Without a forecast interval, it is impossible to determine if an observed value is within an expected range of values. In practical terms, the forecast value and interval work together to help the end user interpret the results of the forecast.

A comparison of the forecast intervals against the holdout values found a disproportionate number (84%) of actual values located in the first forecast interval. This further validates the use of Box-Jenkins as a forecasting tool in public health practice; if the forecasting associated with this study followed a normal distribution, the number of actual values located in the first forecast interval would approximate 68%. In practical terms, forecast intervals may be as important as forecast accuracy as it is possible to have forecasts that are not considered accurate but have a value contained within the first or second forecast interval. There were several examples of this occurring in the analysis. It was also found that when an actual value is located within the second forecast interval it progresses to outbreak status 54% of the time; this suggests that when setting a threshold for action (i.e., increase vigilance, active surveillance, etc.) consideration should be given to setting the threshold at the boundary of the first and second forecast interval.

Jurisdictional Size and Forecast Accuracy

There was no statistical relationship between jurisdictional size and forecast accuracy. Again, this implies that Box-Jenkins forecasting models may be a useful tool for public health practice. It is often assumed that the establishment of baselines is of limited value in small jurisdictions due to the potential volatility associated with small number analysis. It has been demonstrated, through this study that Box-Jenkins forecasting controls for this volatility and produces usable results for public health jurisdictions regardless of size. This is an important finding as nearly 2/3 of all public health jurisdictions in the United States serve populations less than 50,000 and are classified as Small by the National Association of City and County Health Officers.

It is interesting to note that jurisdictional size and rate of disease does not diminish forecast accuracy. The point where the model fails is directly linked to jurisdictional size and rate of disease and is simply a function of population and rate of disease; frequently, diseases with small rates in smaller public health jurisdictions do not occur for several years, if ever. This results in forecasts and forecast intervals of zero. While technically correct, they do not represent useful information

Model Failure Point

The results of this study demonstrate that the failure point of the model is due to the lack of available data and not jurisdictional size or rate of disease. As with the other findings, this result suggests that the Box-Jenkins model is useful for public health practice. As expected, there are statistically significant findings that smaller jurisdictions and diseases with lower rates have a greater number of incomplete forecasts. Incomplete forecasts are a direct function of the population size and rate of disease. For example, in a jurisdiction with a population of

50,000 and a disease rate of 0.2/100K, one would expect to see the disease report for that disease once every 10 years.

Based upon these findings it is apparent that it is possible to determine the cutoff points of useful forecasting based upon jurisdictional size and rate of disease. Using the National Association of City and County Health Officials department size designation of Small, Medium, and Large Table 8 displays the point where this occurs.

Table 8: *Determination of Cutoff Points for Box-Jenkins Forecasting Ability Based Upon Jurisdictional Size and Rate of Disease*

	Large (500,000+)	Medium (50,000–499,999)	Small (< 50,000)
Giardiasis 15.2/100K	Usable Forecasts	Usable Forecasts	Usable Forecasts
Campylobacter 13.6/100K	Usable Forecasts	Usable Forecasts	Usable Forecasts
Salmonella 12.8/100K	Usable Forecasts	Usable Forecasts	Usable Forecasts
Strep pneumoniae 5.7/100K	Usable Forecasts	Use with Caution	No
Coccidioidomycosis 2.3/100K	Usable Forecasts	Use with Caution	No
Lyme 0.9/100K	Usable Forecasts	Use with Caution	No
Encephalitis 0.4/100K*	Use with Caution	No	No
Malaria 0.2/100K	Use with Caution	No	No

Note. The study had four diseases with a rate of 0.4/100K. Encephalitis represents the findings based upon jurisdictional size and rate of disease as all failed to produce usable forecasts at the same point.

Discussion

The results of the study indicate that it is possible to predict future disease burdens within a community based upon the historical surveillance data within local health jurisdictions of varying size. The statistically insignificant results demonstrate that the forecasting model is effective, and the lack of differences between the accuracy of the projections in jurisdictions of varying size and rates of disease demonstrates that forecasting is a successful technique that may have a practical application for local health departments. Smaller jurisdictions may have challenges related to the inability to make projections due to the interaction between jurisdictional size and rate of disease. The results imply that smaller jurisdictions and diseases with lower rates may have more forecasts that cannot be completed due to insufficient data. Beyond these findings, there are practical applications for Box-Jenkins forecasting results as well.

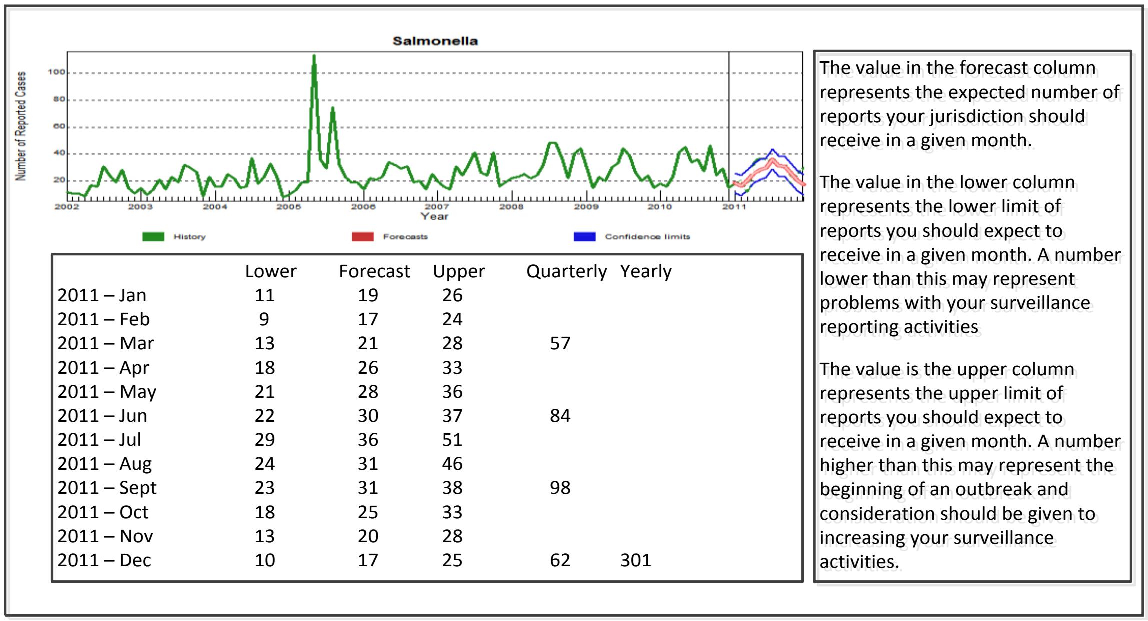
The timely detection, investigation, control, and prevention of outbreaks and other long-term public health problems require a well-trained and competent epidemiology workforce. The National Association of County and City Health Officials assessed the size of the epidemiology workforce in local health departments and reported that 1,500 epidemiologists worked in local health (Novich, 2011). Using data from the same report, the estimated number of local health departments without an epidemiologist is 72%. Many reasons are cited for this low capacity, the most frequent being uncompetitive pay and the lack of trained epidemiologists.

For many health departments, the job classification of epidemiologist falls to other professional disciplines, such as a registered nurse or environmental health specialist (Moehrle, 2008). This reality requires many public health professionals to work in areas that were not part of their formal education, which may limit their understanding on the nuances associated with disease surveillance activities, including when to increase their surveillance and disease control activities based upon an increase in the number of disease reports. Additionally, this lack of formal epidemiological training may make the use of complex statistical models difficult for them to utilize and interpret.

Based upon inexpensive and readily available software, the forecasting techniques used in this study are simple to complete and provide results that are easy to understand. Outputs from the forecast produce a point projection for the number of disease events in a specific timeframe and forecast intervals that help put the projected number into context. This information may be used to evaluate the creditability of an astute observer report and the results of routine data evaluation. A nurse or environmental health specialist can use this information to know when to increase prevention activities or vigilance to a specific disease instead of relying upon past experience or intuition.

Using actual forecast results and the data used to create them a plausible scenario of how Box-Jenkins forecasts may be used in public health is presented. A nurse, acting in the capacity of a local health department's epidemiologist, receives the following series of disease reports.

During the first quarter of 2011, there are 18, 20, and 12 confirmed cases of Salmonella reported to the health department. During the second quarter, the number of confirmed cases more than doubles with 33, 38, and 36 cases reported to the health department. To a registered nurse, these confirmed reports may appear to be the beginning of an outbreak. Using the results of the Box-Jenkins forecasts contained in Figure 6, the nurse compares the number of confirmed cases of Salmonella against the forecast values. While the number of confirmed cases is high when compared to the forecast values, they are within the expected range of the forecast interval. The nurse determines that the number of reports does not represent an outbreak, but based upon the number of confirmed cases being on the high side of the1st forecast interval, sends a



	Lower	Forecast	Upper	Quarterly	Yearly
2011 – Jan	11	19	26		
2011 – Feb	9	17	24		
2011 – Mar	13	21	28	57	
2011 – Apr	18	26	33		
2011 – May	21	28	36		
2011 – Jun	22	30	37	84	
2011 – Jul	29	36	51		
2011 – Aug	24	31	46		
2011 – Sept	23	31	38	98	
2011 – Oct	18	25	33		
2011 – Nov	13	20	28		
2011 – Dec	10	17	25	62	301

The value in the forecast column represents the expected number of reports your jurisdiction should receive in a given month.

The value in the lower column represents the lower limit of reports you should expect to receive in a given month. A number lower than this may represent problems with your surveillance reporting activities

The value is the upper column represents the upper limit of reports you should expect to receive in a given month. A number higher than this may represent the beginning of an outbreak and consideration should be given to increasing your surveillance activities.

Figure 6. Forecast values and upper/lower forecast interval for Salmonella, 2012. Vertical line separates model’s historical values from predicted values. Forecast intervals are the upper and lower horizontal lines on the right of the vertical separator.

newsletter to the medical providers in her community reporting the findings and asking them to consider obtaining stool samples from their patients presenting with symptoms consistent with Salmonella.

Limitations

Some limitations of this study need to be taken into account when interpreting the results. In this study, the data used to create the forecasts were obtained from a NEDSS compliant database. As such, these data are from a passive surveillance system; the possible biases in disease reporting and potential underreporting may influence the precision of this analysis. Inconsistencies in the data were also noted ranging from variations in the naming of disease, changing case definitions, and apparent data gaps associated with specific diseases. These problems resulted in the inability to complete the analyses with six of the 22 diseases that were initially selected for use in this study.

There are known delays between diagnosis and report dates; however, this limitation should improve with the implementation of an electronic-based reporting system. The level of analysis is also a limitation. The unit of analysis was by month and to increase the usefulness of the forecast the granularity needs to be at a finer level.

Using MAPE to analyze the accuracy of the forecasts introduced a level of uncertainty into the absolute accuracy calculations. Fifty-two perfect reports were not included in the overall accuracy calculation as perfect forecast result in a division by zero problems. An alternative method to determine accuracy should be considered; however, the reader is cautioned to be aware of the limitations of each of the accuracy determination methods.

The final limitation is that there is a “rate hole” in the analysis. This hole represents diseases that have a rate between 5.7 – 12.6/100K. The aforementioned data problems are responsible for a portion of this limitation, but the largest contributor is based upon the fact that most of the reportable diseases in Utah do not fall into this range; this limitation impacts the ability to determine the cutoff point of Box-Jenkins forecasting ability based upon rate of disease and jurisdictional size.

Future Studies

Box-Jenkins was chosen for evaluation as it is routinely used to support forecasting problems in business, economic, and control-engineering applications, yet it had not been systematically examined for use with public heath surveillance data. The results of this study show that it is possible to produce Box-Jenkins forecasts that are equal to or greater than 75% accurate. However, these findings open additional questions that should be assessed in future studies. For example, what other types of analytical techniques are there that can produce forecasts as accurate as or more accurate than Box-Jenkins? Do forecasts based upon exponential smoothing models (i.e., Simple, Holt, and Winters) produce better forecast models than Box-Jenkins? Is there a role for simple moving averages, and how does one support forecasting models that have a disproportionate number of zeros in the historical dataset? What role, if any, is there for Statistical Process Control and Shewart Control Charts? A future study that answers these questions could produce results that will allow local health jurisdictions to apply the appropriate statistical model to create a baseline that is disease specific and based upon rate of disease and jurisdictional size. Moreover, local, state, and federal public health jurisdictions should consider the use of multiple systems and indicators, including forecasting, to monitor the ongoing health of the communities they serve. Each system has its own unique set of strengths and weakness, and no single system will answer the underlying questions about when resources, including investigations, are expended. This question is especially important in smaller jurisdictions with limited manpower and budgets, again, strongly suggesting the use of multiple systems.

Finally, a future study needs to be conducted across a larger dataset with a different pattern of disease burden in the population to examine how Box-Jenkins forecasts perform in the aforementioned “rate hole” that encompasses diseases with rates between 5.7 – 12.6/100K. Not only will this address model performance but will help determine the cutoff points for model usage based upon rate of disease and jurisdictional size.

Conclusion

The results of this study showed that the cumulative accuracy of all the forecast trials was 71%. Although this is lower than the 75% threshold associated with the original research question, it was found that 63% of the disease specific forecasts had accuracies greater than 75%. Thus, it is valid to state that Box-Jenkins forecasts can produce disease specific forecasts that are equal to or greater than 75% accurate, but not for all diseases of public health significance.

The lack of a statistical correlation between the rate of disease and forecast accuracy as well as jurisdictional size and forecast accuracy demonstrated that Box-Jenkins forecasts retain their accuracy regardless of rate of disease or jurisdictional size. This is important as the majority of health departments in the United States serve populations < 50,000. However, it was noted that smaller health departments and diseases with low rates tend to be statistically more likely to have a higher number of incomplete forecasts due to the lack of data necessary to create a forecast.

A comparison of the forecast intervals against the holdout values found a disproportionate number of actual values located in the first forecast interval. This further validates the use of Box-Jenkins as a tool for public health practice. In practical terms, forecast intervals may be as important as forecast accuracy as it is possible to have forecasts that are not considered accurate but have a value contained within the first or second forecast interval. There were several examples of this occurring in the analysis.

An important component of any epidemiologically based analysis is the ability to rapidly identify the difference between the expected disease burden within specific populations and a disease burden that actually represent an abnormal finding or the beginning of a disease outbreak or epidemic. The use of Box-Jenkins time series models to create forecasts of these expected disease burdens may be a useful tool to support these types of analysis. An additional benefit is that these forecasts appear to be useful in jurisdictions of varying size. Finally, the use of Box-Jenkins as a tool should encourage epidemiologist to look beyond the traditional biostatistical methods associated with disease surveillance, as the underlying methodology associated with its use are borrowed from the unlikely discipline of business management and are based upon forecasting future sales cycles, not disease. The use of the techniques described in this study can easily be automated and show promise as an effective tool to assist local public health jurisdictions in their efforts to monitor and control disease.

REFERENCES

Armstrong, J. S. (2001). *Principles of forecasting: A handbook for researchers and practitioners*. Boston: Kluwer Academic.

Armstrong, J. S., & Collopy, F. (1994). Error measures for generalizing about forecasting methods: Empirical comparison. *International Journal of Forecasting*. *8*, 69-80.

Baer, A., Rodriguez, C. V., & Duchin, J. S. (2011). An automated system for public health surveillance of school absenteeism. *Journal of Public Health Management and Practice, 17*(1), 59-64.

Berube, M. S. (Ed.). (1985). *American heritage dictionary* (2nd ed.). Boston, MA: Houghton Mifflin.

Box, G. E., & Jenkins, G. M. (1994). *Time series analysis: Forecasting and control* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Box, G. E., Jenkins, G. M., & Bacon, D. W. (1967). Models for forecasting seasonal and nonseasonal time series. In B. Harris (Ed.), *Spectral analysis of time series*. New York, NY: John Wiley & Sons.

Boylan, J. (2005). Intermittent and lumpy demand: A forecasting challenge. *The International Journal of Applied Forecasting, 1*, 36-42.

Buckeridge, D. (2010). *A method for evaluating outbreak detection in public health surveillance systems that use administrative data* (Unpublished doctoral dissertation). Stanford University, CA.

Caldwell, J. G. (n.d.) *The Box-Jenkins forecasting technique.* Retrieved March 3, 2012, from http://www.foundationwebsite.org/BoxJenkins.htm

Centers for Disease Control and Prevention. (n.d.). *Excellence in curriculum innovation through teaching epidemiology and the science of public health.* Retrieved September, 17, 2012, from http://www.cdc.gov/excite/classroom/outbreak/objectives.htm

Centers for Disease Control and Prevention. (1997). Case definitions for infectious conditions under public health surveillance. *Morbidity and Mortality Weekly Report 46*(RR10), 1-55.

Centers for Disease Control and Prevention. (2006). Notice to readers: Changes in presentation of data from the national notifiable diseases surveillance system—January 13, 2006. *Morbidity and Mortality Weekly Report, 55*(01).

Centers for Disease Control and Prevention. (2008a). *CDC solutions, national electronic disease surveillance system.* Retrieved October 20, 2011, from http://www.cdc.gov/phin/library/documents/pdf/111759_NEDSS.pdf

Centers for Disease Control and Prevention. (2008b). *FY 2008 annual performance report.* Atlanta, GA: Author.

Centers for Disease Control and Prevention. (2009). Summary of notifiable diseases—United States, 2009. *Morbidity and Mortality Weekly Report, 58*(53), 1-100.

Centers for Disease Control and Prevention. (2011). Public health now and then: Celebrating 50 years of MMWR at CDC. *Morbidity and Mortality Weekly Report, 60* (Supplement), 1-125.

Centers for Disease Control and Prevention. (2012). *National notifiable disease surveillance system: Monitoring the spread of diseases.* Retrieved March 3, 2012, from http://www.cdc.gov/osels/ph_surveillance/nndss/nndsshis.htm

Chen, H., Zeng, D., & Yan, P. (2009). *Infectious disease informatics: Syndromic surveillance for public health and bio-defense* (1st ed.). New York, NY: Springer.

Council of State and Territorial Epidemiologists. (2010). *NEDSS assessment report*. Atlanta, GA: Council of State and Territorial Epidemiologists.

Council of State and Territorial Epidemiologists. (2012). *CSTE position statement number: 10-ID-07.* Retrieved February 20, 2011, from http://www.cdc.gov/osels/ph_surveillance/nndss/casedef/hepatitisacurrent.htm

Cristofferson, P. F. (1998). Evaluating forecast intervals. *International Economic Review, 37*(4), 841-862.

Dalton, C. B., Haddix, A., Hoffman, R. E., & Mast, E. E. (1996). The cost of a food-borne outbreak of hepatitis A in Denver, Colorado. *Archives of Internal Medicine, 156*(9), 1013-1016.

Forecast PRO XE (Version 5.5) [Computer Software]. Belmont, MA: Business Forecast Systems.

Gesteland, P. H., Wagner, M. M., Chapman, W. W., Espino, J. U., Tsui, F. C., Gardner, R. M., & Haug, P. J. (2002). Rapid deployment of an electronic disease surveillance system in the state of Utah for the 2002 Olympic

Winter Games. *Proceedings AMIA Annual Symposium. AMIA Symposium,* 285-289.

Geurts, M. D., & Ibrahim, I. B. (1975). Comparing the Box-Jenkins approach with the exponentially smoothed forecasting model application to Hawaii tourists. *Journal of Marketing Research, 12*(2), 52-66.

Goldstein, B. D. (2010). *Biowatch and public health surveillance evaluating systems for the early detection of biological threats.* Washington, DC: National Academy of Sciences.

Henning, K. J. (2004). Overview of syndromic surveillance: What is syndromic surveillance? *Morbidity and Mortality Weekly Report 53*(Supplement), 5-11.

Helfenstein U. (1986) Box-Jenkins modelling of some viral infectious diseases. *Statistics in Medicine, 5,* 37-47.

Helfenstein U. (1996) Box-Jenkins modelling in medical research. *Statistical Methods in Medical Research.5,* 3.

Heymann, D. L. (Ed.). (2008). *Control of communicable diseases manual* (17th ed.). Washington, DC: American Public Health Association.

Hogarth, R. M., & Makridakis, S. (1981). Forecasting and planning: An evaluation. *Management Science, 27*(2), 115-138.

Hutwagner, L. C., Maloney, E. K., Bean, N. H., Slutsker, L., & Martin, S. M. (1997). Using laboratory-based surveillance data for prevention: An algorithm for detecting salmonella outbreaks. *Emerging Infectious Diseases, 3*(3), 395-400.

Hyndman, R. J. (2006). Another look at forecast demand-accuracy metrics for intermittent demand. *The International Journal of Applied Forecasting, 4*, 43-46.

Jantch, E. (1967). *Technological forecasting in perspective.* Paris, France: OECD.

Kleinman, K., Abrams, A., Yih, W. K., Platt, R., & Kulldorff, M. (2006). Evaluating spatial surveillance: Detection of known outbreaks in real data. *Statistics in Medicine, 25*(5), 755-769.

Last, J. (Ed.). (1995). *A dictionary of epidemiology* (3rd ed.). New York, NY: Oxford University Press.

Leep, C., J. (2007). *The local health department workforce: Findings from the 2005 national profile of local health departments study.* Washington, DC: National Association of County & City Health Officials.

Levenback, H., & Cleary, J. P. (2006). *Forecasting practice and process for demand management.* Belmont, CA: Thomson Brooks/Cole.

Lipsitch, M., & Viboud, C. (2009). Influenza seasonality: Lifting the fog. *Proceedings of the National Academy of Sciences of the United States of America, 106*(10), 3645-3646.

Makridakis, S., Chatfield, M., Hilbon, M. J., Lawrence, T., Mills, K., Ord, K., & Simmons, L. F. (1993). The M2-competion: A real time judgmentally based forecasting study. *International Journal of Forecasting, 9*, 5-22.

Makridakis, S., & Wheelwright, S. C. (Eds.). (1987). *The handbook of forecasting. A manager's guide* (2nd ed.). New York, NY: John Wiley & Sons.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2000). *Forecasting methods and applications* (3rd ed.). New York, NY: John Wiley & Sons.

Mentzer, J. T., & Cox, J. E. (1984). Familiarity, application and performance of sales forecasting techniques. *Journal of Forecasting, 3*(1), 27-36.

Mikanatha, N. M. (Ed.). (2007). *Infectious disease surveillance* (2nd ed.). New York, NY: John Wiley & Sons.

Moehrle, C. (2008). Who conducts epidemiology activities in local public health departments? *Public Health Reports, 123*(1), 6-7.

Morris, G., Snider, D., & Katz, M. (1996). Integrating public health information and surveillance systems. *Journal of Public Health Management and Practice, 2*(4), 24.

National Office of Vital Statistics. (1953). Reported incidence of selected notifiable disease: United States, each division and state 1925-50. (Report No. 37). *Vital Statistics Special Reports (National Summaries), 37*, 1180-1181. Washington, DC: Government Printing Office.

Novich, R. (2011). *2010 national profile of national health departments.* Washington, DC: National Association of City and County Health Officials.

Overhage, M. J., Grannis, S., & McDonald, C. J. (2008). A comparison of the completeness and timeliness of automated electronic laboratory reporting and spontaneous reporting of notifiable conditions. *American Journal of Public Health, 98*(2), 344-350.

Pankratz, A. (1983). *Forecasting with univariate Box-Jenkins concepts and cases*. New York, NY: John Wiley & Sons.

RefWorks (Version 2.0) [Computer Software]. New York, NY: Cambridge Information Group.

Restaurant industry weighs vaccination for food handlers. (1997, Winter). *National Restaurant Association, 1*(4). Retrieved May 20, 2012, from http://www.hepatitiscontrolreport.com/vol/v1n41.html

Rycroft, R. S. (1995). Student editions of forecasting software: A survey. *International Journal of Forecasting, 11*, 337-351.

SPSS statistics premium graduate student version (Version 21.0) [Computer Software]. Armonk, NY: IBM.

Stellwagen, E., & Goodrich, R. (2008). *Forecast pro user's guide* (5th ed.). Belmont, MA: Business Forecast Systems.

Stroup, D. F. (1989). Detection of aberrations in the occurrence of notifiable diseases surveillance data. *Statistics in Medicine, 8*(3), 323-329.

Stroup, D. F., Wharton, M., Kafadar, K., & Dean, A. G. (1993). Evaluation of a method for detecting aberrations in public health surveillance data. *American Journal of Epidemiology, 137*(3), 373-380.

Teutsch, S. M., & Churchill, R. E. (Eds.). (1994). *Principles and practice of public health surveillance* (1st ed.). New York, NY: Oxford University Press.

Unkel, S., Farrington, P., Garthwaite, P., Robertson, C., & Andrews, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: A review. *Journal of the Royal Statistical Society, 75*(1), 49-82.

U.S. Census Bureau. (2012). *State and county quickfacts: UT.* Retrieved January 15, 2012, from http://www.census.gov/

Utah administrative code. R386-702-3 Communicable disease rule. Reportable diseases, emergency illnesses, and health conditions. R386. (2012). Utah Department of Administrative Services, Division of Administrative Rules.

Utah Department of Health. (2011). *Utah health status update: Health innovation summit.* Presented at the Health Innovation Summit, Salt Lake City, UT.

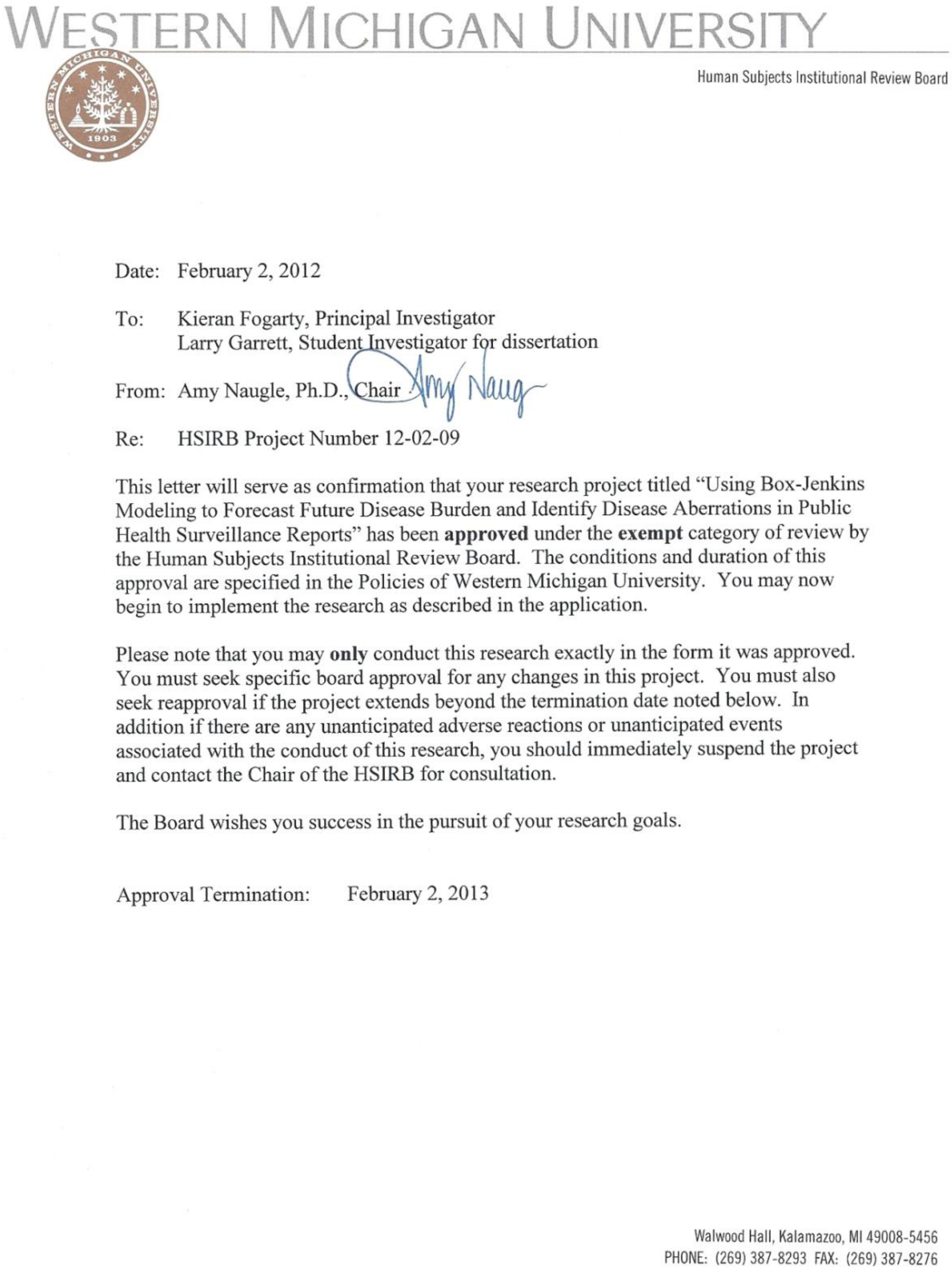
Uziel, A., & Stone, L. (2012). Determinants of periodicity in seasonally driven epidemics. *Journal of Theoretical Biology, 305*, 88-95.

Williamson, G. D., & Weatherby, H. G. (1999). A monitoring system for detecting aberrations in public health surveillance reports. *Statistics in Medicine, 18*(23), 3283-3298.

Winklhofer, H., Diamantopoulos, A., & Witt, S. F. (1996). Forecasting practice: A review of the empirical literature and an agenda for future research. *International Journal of Forecasting, 12*, 193-221.

Appendix A

Human Subjects Institutional Review Board Letter of Approval



WESTERN MICHIGAN UNIVERSITY
Human Subjects Institutional Review Board

Date: February 2, 2012

To: Kieran Fogarty, Principal Investigator
Larry Garrett, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Chair Amy Naug

Re: HSIRB Project Number 12-02-09

This letter will serve as confirmation that your research project titled "Using Box-Jenkins Modeling to Forecast Future Disease Burden and Identify Disease Aberrations in Public Health Surveillance Reports" has been **approved** under the **exempt** category of review by the Human Subjects Institutional Review Board. The conditions and duration of this approval are specified in the Policies of Western Michigan University. You may now begin to implement the research as described in the application.

Please note that you may **only** conduct this research exactly in the form it was approved. You must seek specific board approval for any changes in this project. You must also seek reapproval if the project extends beyond the termination date noted below. In addition if there are any unanticipated adverse reactions or unanticipated events associated with the conduct of this research, you should immediately suspend the project and contact the Chair of the HSIRB for consultation.

The Board wishes you success in the pursuit of your research goals.

Approval Termination: February 2, 2013

Walwood Hall, Kalamazoo, MI 49008-5456
PHONE: (269) 387-8293 FAX: (269) 387-8276

Appendix B

List of Acronyms

List of Acronyms

ARMA: Autoregressive Moving Average

ARIMA: Autoregressive Integrated Moving Average

CSTE: Council of State and Territorial Epidemiologists

CUSUM: Cumulative Sum Chart

CDC: The Centers for Disease Control and Prevention

ELR: Electronic Laboratory Report

ESP: Electronic Surveillance Project

MAD: Mean Absolute Deviation

MSE: Mean Squared Error

MAPE: Mean Absolute Percentage Error

MMWR: Morbidity and Mortality Weekly Report

NBS: NEDSS Base System

NETSS: National Electronic Telephonic System for Surveillance

NEDSS: National Electronic Disease Surveillance System

NNDSS: National Notifiable Disease Surveillance System

PHS: Public Health Service

SMA: Simple Moving Average

SPC: Statistical Process Control

UDOH: Utah Department of Health

YTD: Year to Date

Appendix C

National Reportable Diseases

Nationally Notifiable Disease Listing:

The following conditions are immediately notifiable to CDC for all or some cases, as specified in the relevant CSTE position statements:

1. Influenza, novel
2. Measles
3. Plague Anthrax
4. Botulism
5. Brucellosis
6. Diphtheria
7. Polio
8. Rabies in an animal
9. Rabies in a human
10. Rubella
11. SARS Coronavirus
12. Smallpox
13. Tularemia
14. Viral hemorrhagic fevers
15. Yellow fever
16. Arboviral Disease, Calif. Serogroup
17. Arboviral Disease, Eastern equine
18. Arboviral Disease, Powassan
19. Arboviral Disease, St. Louis
20. Arboviral Disease, West Nile Virus

21. Arboviral Disease, Western equine
22. Cancer
23. Chancroid
24. Chlamydia trachomatis
25. Cryptosporidiosis
26. Cyclosporiasis
27. Dengue Virus Infections
28. Ehrlichiosis/Anaplasmosis
29. Escherichia coli, Shiga toxin-producing (STEC)
30. Giardiasis
31. Gonorrhea
32. Haemophilus influenzae
33. Hantavirus pulmonary syndrome
34. Hemolytic Uremic Syndrome, Post-diarrheal
35. Hepatitis A
36. Hepatitis C
37. HIV
38. Influenza-associated mortality, pediatric
39. Lead, exposure screening test result
40. Legionellosis
41. Listeriosis
42. Malaria
43. Meningococcal Disease (Neisseria meningitidis)

44. Mumps
45. Pertussis
46. Pesticide-related Illness
47. Psittacosis
48. Q fever
49. Spotted Fever Rickettsiosis
50. Rubella, Congenital Syndrome
51. Salmonellosis
52. Shigellosis
53. Silicosis
54. Staphylococcus (VISA, VRSA)
55. Streptococcus (STSS, IPD)
56. Syphilis
57. Tetanus
58. Trichinellosis (Trichinosis)
59. Tuberculosis
60. Typhoid fever
61. Varicella (Chickenpox)
62. Vibrio cholera
63. Vibriosis
64. Waterborne Disease Outbreak

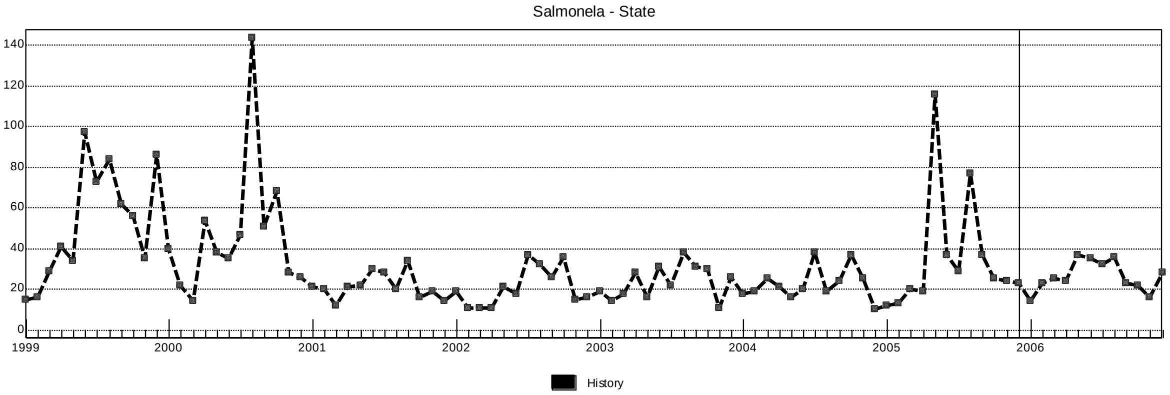
The following conditions are provisionally included in the Nationally Notifiable Condition List. Conditions included on a provisional basis are those which have not yet met the criteria outlined in CSTE position statement 08-EC-02; for example, a CDC Case Notification Request may be incomplete, or a revised position statement containing a case reporting definition may be incomplete.

65. Foodborne outbreaks
66. Hansen's disease (leprosy)
67. Hepatitis B, acute
68. Hepatitis, viral, chronic: Hepatitis B
69. Lyme disease
70. Toxic-shock syndrome (non-streptococcal)

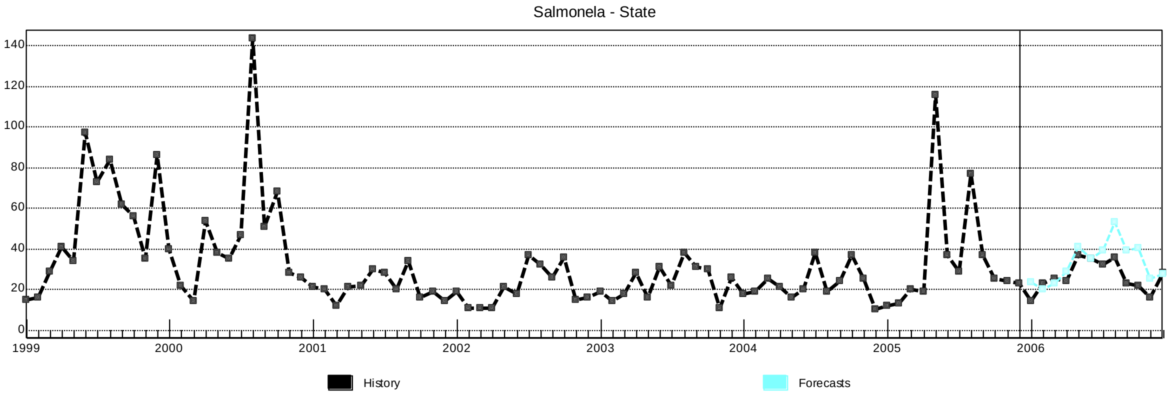
Appendix D

Outlier Identification and Correction

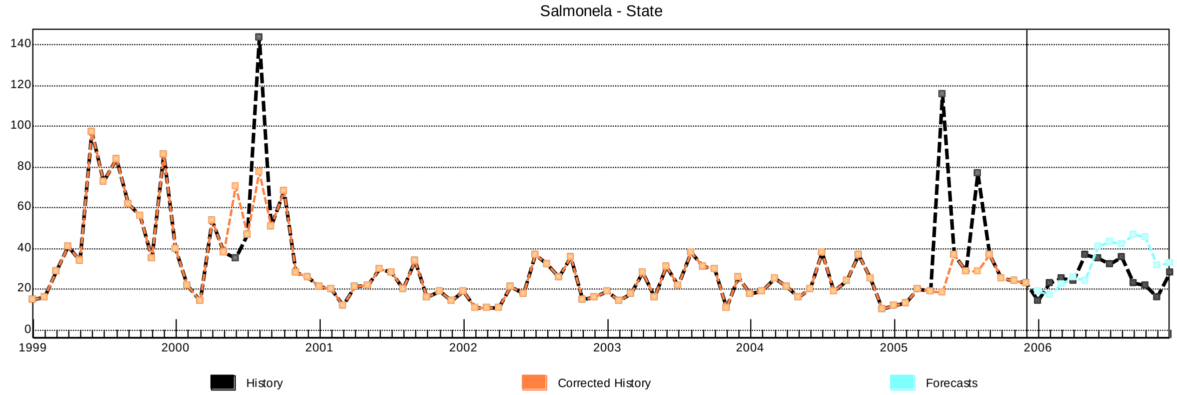
This appendix shows graphically the effect of the outlier identification and correction process on an actual forecast. The data used in this example represents statewide disease reports of Salmonella for the years 1999-2006. A holdout sample has been applied to year 2006 and is represented by the vertical line.



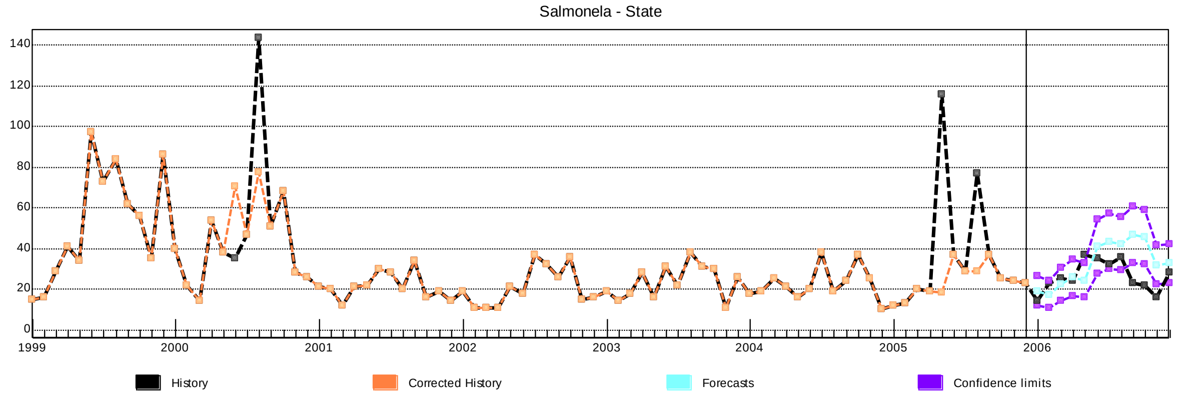
This graph shows the historical record for cases of reported cases of Salmonella for the years 1999-2006.



The graph shows both the forecast value and the actual value in the holdout sample (*i.e.,* year 2006). No outlier identification or correction has been completed.



The graph shows both the outlier and corrected values. Notice the effect of the correction on the forecast in the holdout sample.



The graph shows the confidence interval associated with the forecast contained in the holdout sample. In this case, all of the forecast values are within the confidence interval.

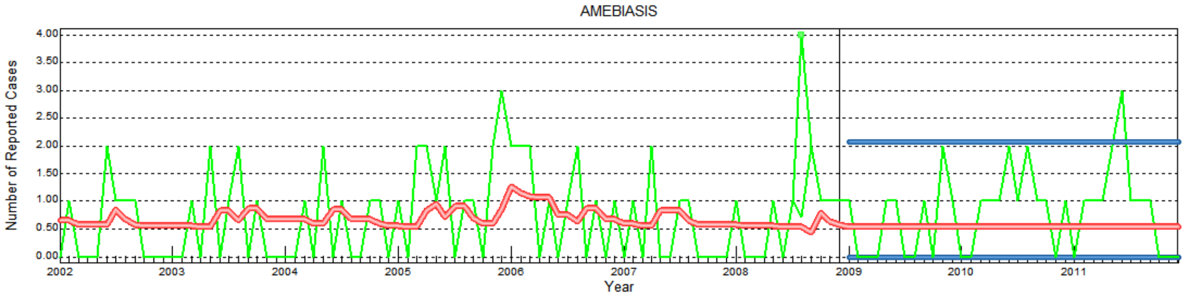
Appendix E

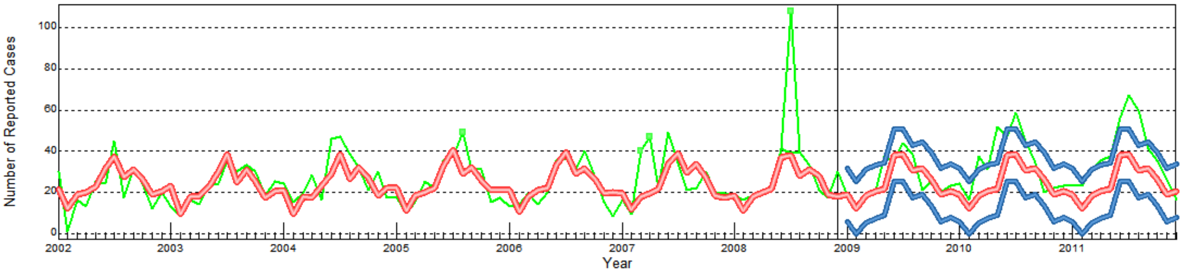
Forecast Results Charts

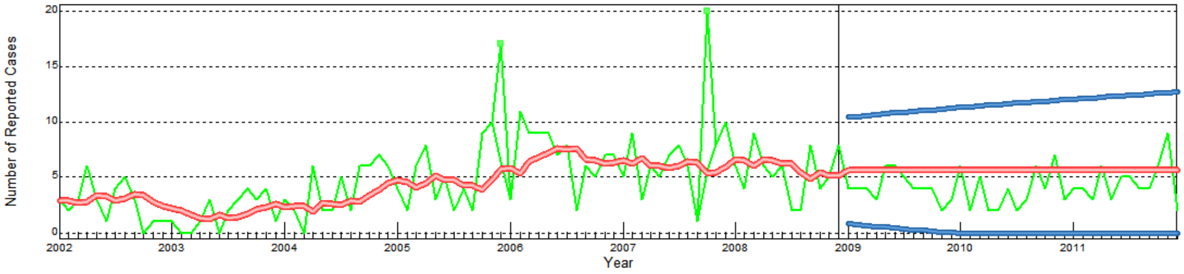
The following graphs show the results of a single three year projection based upon data from the years 2002–2008. They are intended to provide a visual representation of the forecasting process. The forecasting projections associated with this study completed three individual forecasts for each disease. Yearly forecasts result in better overall accuracy. The second forecast interval is displayed.

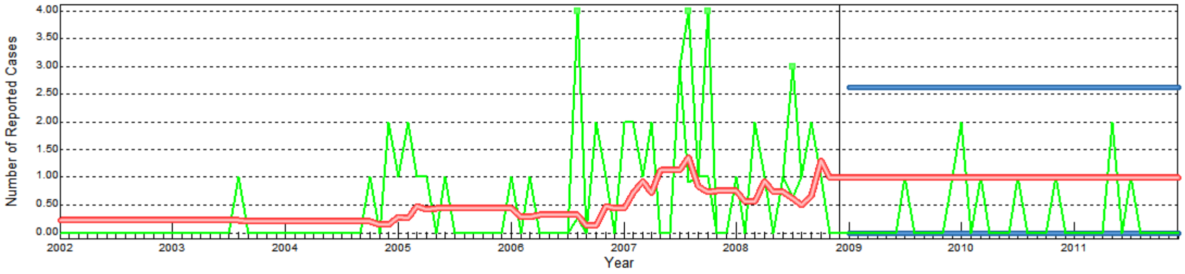
The vertical line on the graph separates the model's historical values from predicted values. The thin line with the most vertical movement on the left of the vertical separator is the historical data. The thick smoothed line on the left of the vertical separator is the fitted value upon which the actual forecast is based. The thin line on the right of the vertical separator is the actual or observed values. The thick smooth line represents the forecast. The forecast intervals are the upper and lower horizontal lines on the right of the vertical separator.

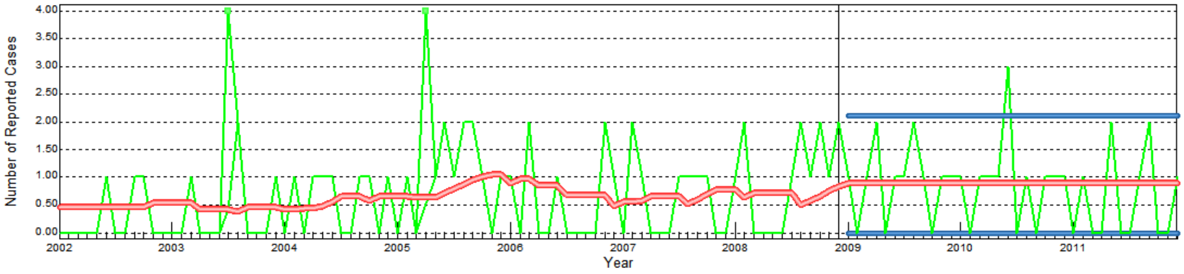
Forecast intervals that rapidly expand represent diseases that benefit from forecasts created on a smaller time frame (i.e., one a year instead of a forecast for three years).

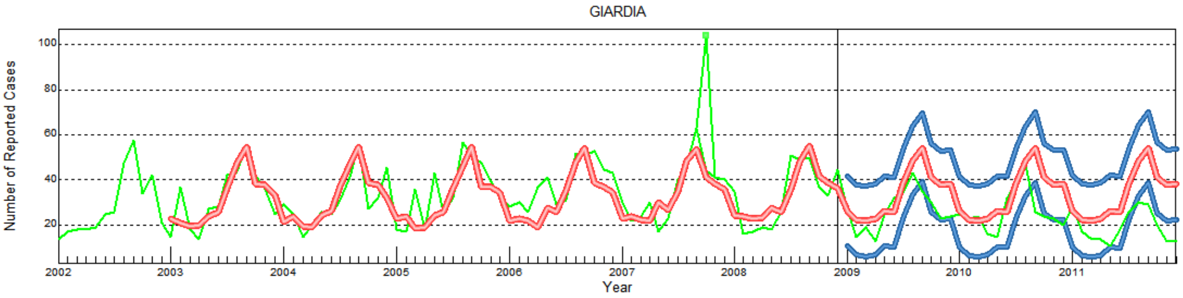


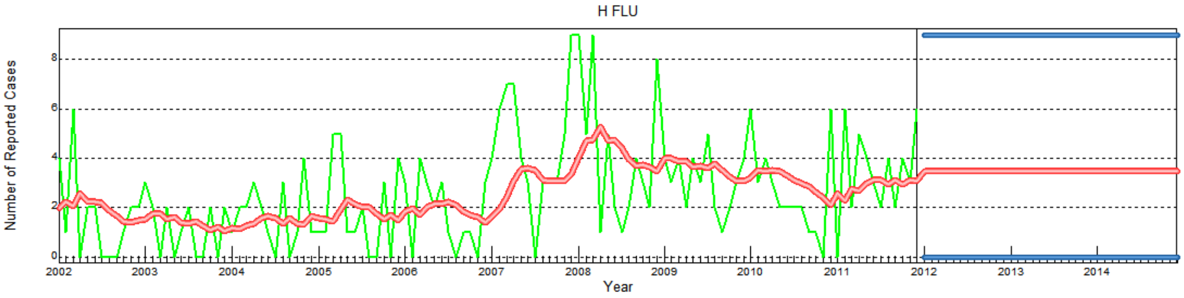

CAMPYLOBACTER

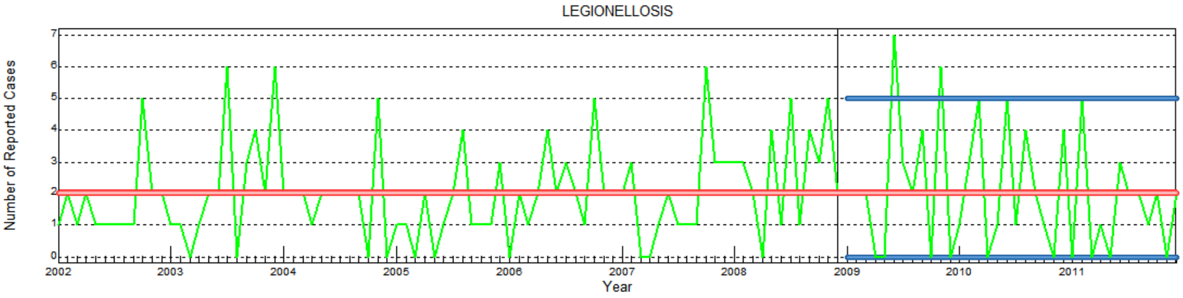

COCCIDIODOMYCOSIS

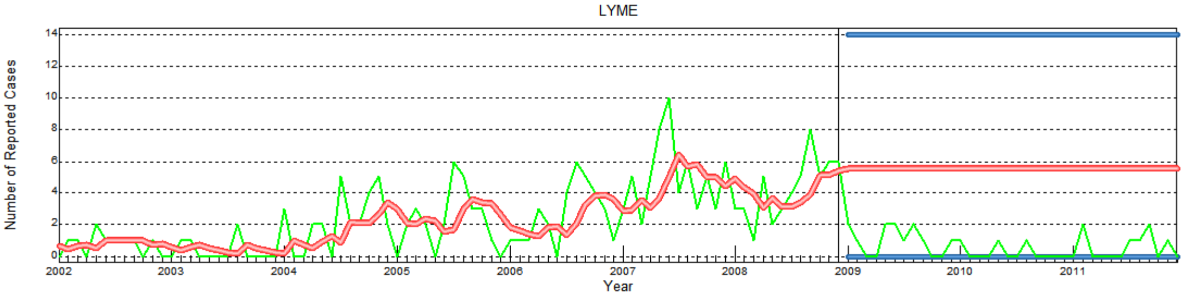

DENGUE FEVER

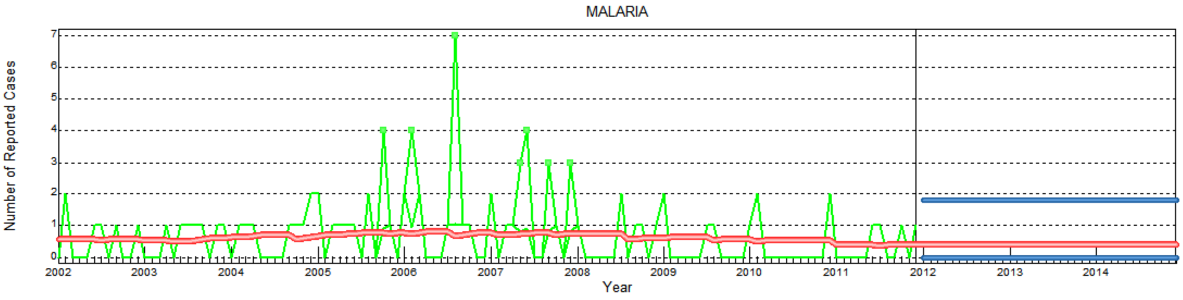

ENCEPHALITIS

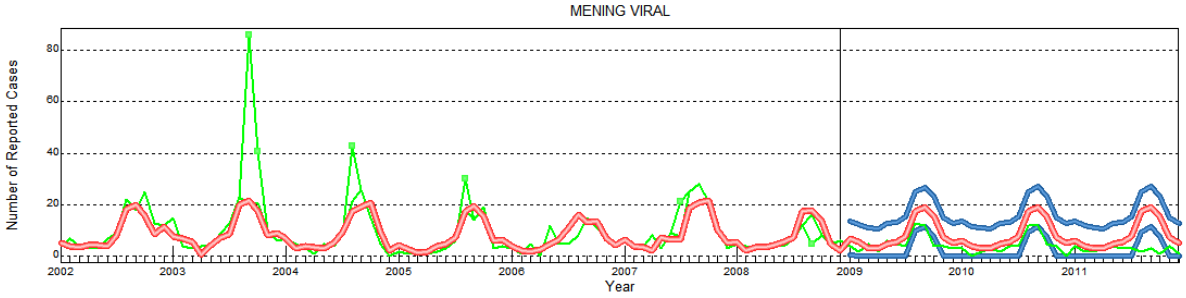

GIARDIA

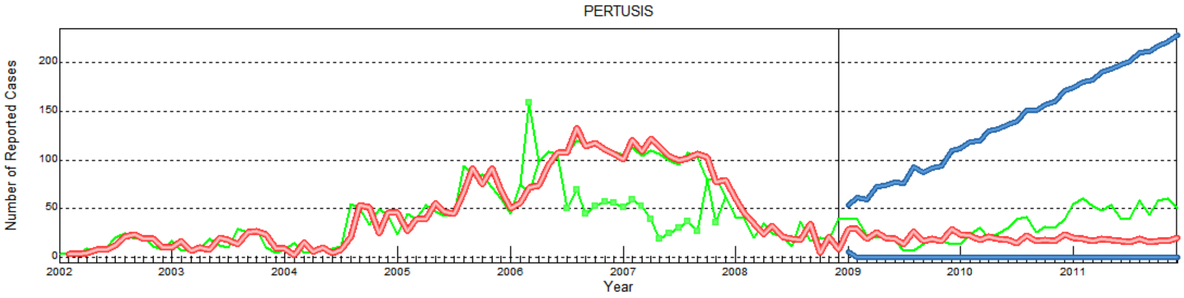

H FLU

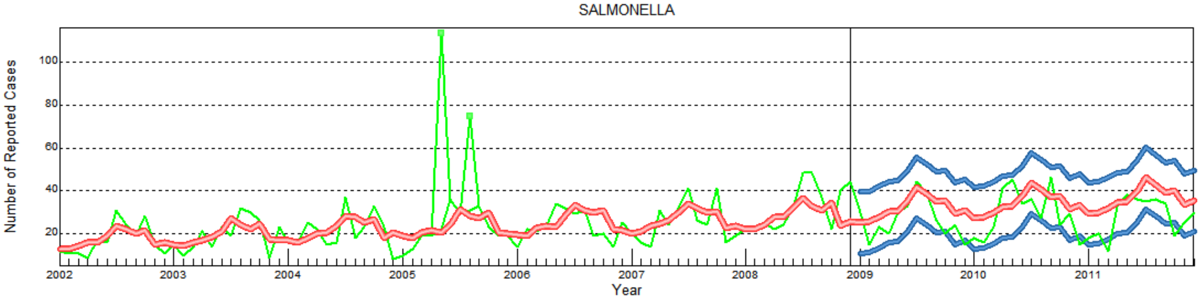

LEGIONELLOSIS

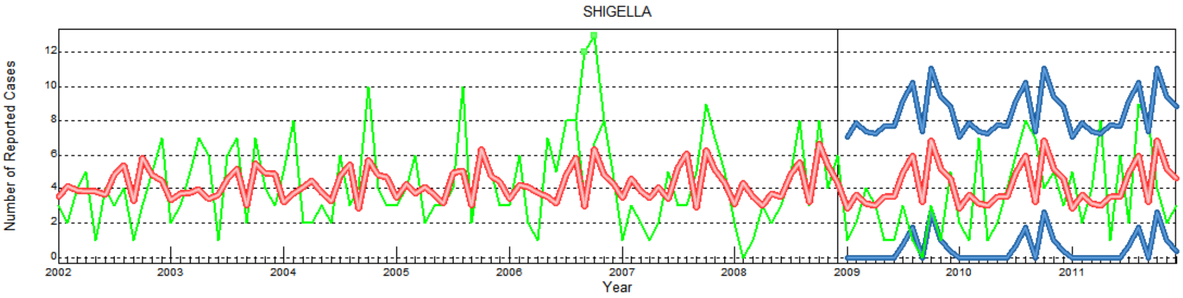

LYME

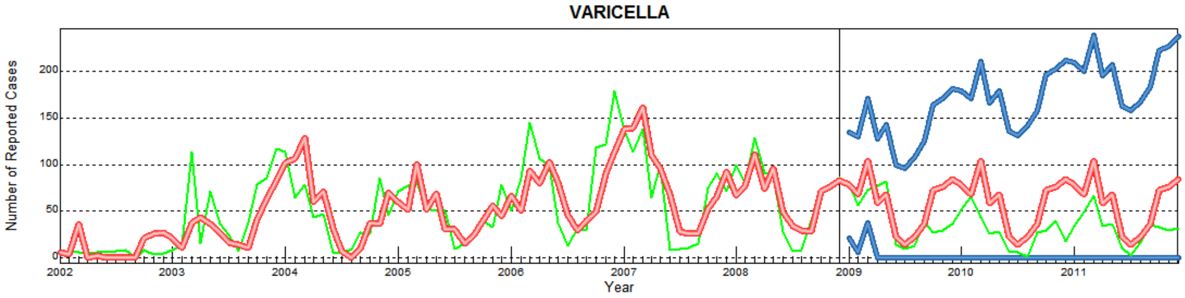

MALARIA
Number of Reported Cases
0
1
2
3
4
5
6
7
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
Year


MENING VIRAL
Number of Reported Cases
0
20
40
60
80
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Year


PERTUSIS
Number of Reported Cases
0
50
100
150
200
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Year


SALMONELLA
Number of Reported Cases
20
40
60
80
100
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Year


SHIGELLA
Number of Reported Cases
0
2
4
6
8
10
12
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Year


VARICELLA
Number of Reported Cases
0
50
100
150
200
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Year