2010

Dynamic Model Pooling Methodology for Improving Aberration Detection Algorithms

Brenton J. Sellati
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/theses

Part of the Agricultural and Resource Economics Commons, Applied Statistics Commons, Longitudinal Data Analysis and Time Series Commons, Other Statistics and Probability Commons, Statistical Methodology Commons, and the Statistical Models Commons

Retrieved from https://scholarworks.umass.edu/theses/535

This thesis is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Masters Theses 1911 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
DYNAMIC MODEL POOLING METHODOLOGY FOR IMPROVING ABERRATION DETECTION ALGORITHMS

A Thesis Presented

by

BRENTON SELLATI

Submitted to the Graduate School of the University of Massachusetts Amherst in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

September 2010
Department of Resource Economics
DEDICATION

This thesis is dedicated to all of the teachers throughout the world who challenge their students to push themselves harder than they thought themselves capable of. Without the influence of such people many of us would not be where we are today.
ACKNOWLEDGEMENTS

Even as I write this I find it difficult to grasp the reality of the situation I currently find myself in; finalizing a thesis which will hopefully allow me to attain a master’s degree in Resource-Economics. When I came to UMass Amherst as an undergraduate, I decided to make Japanese language and literature my major for two reasons—I hated math and had no idea what I wanted to do, other than that I wanted it to be something not involving math. As I write this, I stand on the precipice of embarking upon a career based on mathematical analysis, and I could not be more thrilled and excited for the future. This change would never have been possible without the time, effort, and patience invested in me by my teachers, family, and friends. I’ve learned many lessons from many people over the years, but there are some to whom I especially owe my eternal gratitude.

Jun Ono, my Japanese teacher for four years, for pushing me harder than I ever thought myself capable of. I learned as much about Japanese in her classroom as I did about self-discipline, time management, and striving for perfection. She set the standard of effort she expected from her students by the effort she put in herself. It was under her tutelage that I finally developed a respectable work ethic and for that I will always be grateful.

Marty Holman, another one of my Japanese teachers, without whose influence I don’t know where I’d be today. After struggling with Japanese at the beginning of my sophomore year, I bumped in to Holman-sensei as I was walking out of Whitmore administration building, having just submitted my paperwork to withdraw from my intensive Japanese course. Because of this chance meeting and his offer to vouch for me, instead of quitting Japanese, I ended up studying abroad in Japan the following semester, which became a life changing experience that truly awakened my intellectual curiosity.
Richard Rogers, for teaching me stats really is fun. I approached Rich about doing an independent study in January 2007 after having taken Introductory Business Statistics with him the previous Fall semester. Despite having an incredibly busy schedule already, being the head of undergraduate academic technology for the entire university, Rich took time out of his schedule every week to meet with me and talk about my results. It was through this process that I came to appreciate the true power of statistics because I was able to actually test the validity of my theories using basic data analysis. I became acutely aware of my limited statistical knowledge when I found myself unable to answer many of the lingering questions I had at the time, and this epiphany continues to play a significant role in my desire to continue learning new statistical techniques. As they say, the more you learn the more you realize what you don’t know.

Joe Moffitt, for taking an interest in me when I was a bright-eyed and eager student in his Quantitative Methods for Management Decision Making course. I was originally supposed to work in Japan after finishing my undergraduate degree, but due to a bizarre chain of events, I found out a month before I graduated that I would in fact not be moving to Japan. It was Joe who suggested I consider applying to the Resource-Economics graduate program, which in hindsight was a stroke of serendipity without which I don’t know where I’d be today. When Joe originally asked me to work with him on this thesis, I told him that I was a “big picture” guy, and not go expect too much in the way of mathematical rigor from me. Somewhere along the way that agreement fell by the wayside, and I wouldn’t be able to write this today without Joe’s continued support, timely advice, and belief in my abilities that empowered me to believe I could really pull this off.
I would like to thank all of the other Resource Economics faculty whose courses I had the pleasure of taking throughout my graduate career; in order of appearance, John Spraggon, Tom Stevens, John Stranlund, Dan Lass, Christian Rojas, and Bernie Morzuch. It was often a grueling effort (in the best possible way) to be your student, and looking back I am proud of the progress that each of you pushed me to make in order to succeed in your courses.

I’d like to sincerely thank the USDA/ERS/PREISM under Cooperative Agreement 58-7000-8-0095 and Massachusetts Agricultural Experiment Station Project No. MAS00964 for support and funding of this research. Without the generosity and commitment to improving America of these collaborating agencies, this work would not have been possible.

Finally, I would like to thank my family, who always showed me unwavering love and support throughout all the various stages of my personal evolution, especially when it was unclear how things would turn out. I didn’t always make it easy, but their faith in me never wavered, and I hope that by completing this final requirement for attaining my master’s degree, I will have shown their trust in letting me find my own way well placed.
SYNDROMIC SURVEILLANCE IS DEFINED GENERALLY AS THE COLLECTION AND STATISTICAL ANALYSIS OF DATA WHICH ARE BELIEVED TO BE LEADING INDICATORS FOR THE PRESENCE OF DELETERIOUS ACTIVITIES DEVELOPING WITHIN A SYSTEM. CONCEPTUALLY, SYNDROMIC SURVEILLANCE CAN BE APPLIED TO ANY DISCIPLINE IN WHICH IT IS IMPORTANT TO KNOW WHEN EXTERNAL INFLUENCES MANIFEST THEMSELVES IN A SYSTEM BY FORCING IT TO DEPART FROM ITS BASELINE. COMPARING SYNDROMIC SURVEILLANCE SYSTEMS HAVE LED TO MIXED RESULTS, WHERE MODELS THAT DOMINATE IN ONE PERFORMANCE METRIC ARE OFTEN SORELY DEFICIENT IN ANOTHER. THIS RESULTS IN A ZERO-SUM TRADE-OFF WHERE ONE PERFORMANCE METRIC MUST BE AFFORDED GREATER IMPORTANCE FOR A DECISION TO BE MADE. THIS THESIS PRESENTS A DYNAMIC POOLING TECHNIQUE WHICH Allows FOR THE COMBINATION OF COMPETING SYNDROMIC SURVEILLANCE MODELS IN SUCH A WAY THAT THE RESULTING DETECTION ALGORITHM OFFERS A SUPERIOR COMBINATION OF SENSITIVITY AND SPECIFICITY, TWO OF THE KEY MODEL METRICS, THAN ANY OF THE MODELS INDIVIDUALLY. WE THEN APPLY THIS METHODOLOGY TO A SIMULATED DATA SET IN THE CONTEXT OF DETECTING OUTBREAKS OF DISEASE IN AN ANIMAL POPULATION. WE FIND THAT THIS DYNAMIC POOLING METHODOLOGY IS ROBUST IN THE SENSE THAT IT IS CAPABLE OF SUPERIOR OVERALL PERFORMANCE WITH RESPECT TO SENSITIVITY, SPECIFICITY, AND MEAN TIME TO DETECTION UNDER VARYING CONDITIONS OF BASELINE DATA BEHAVIOR, E.G. CONTROLLING FOR THE PRESENCE OF
or absence of various levels of trend and seasonality, as well as in simulated out-of-sample performance tests.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>v</td>
</tr>
<tr>
<td>Abstract</td>
<td>viii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiii</td>
</tr>
<tr>
<td><strong>CHAPTER</strong></td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION TO SYNDROMIC SURVEILLANCE</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Why syndromic surveillance?</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Historical Instances of Epizootics</td>
<td>4</td>
</tr>
<tr>
<td>1.4 A Dangerous Melting Pot</td>
<td>6</td>
</tr>
<tr>
<td>2 LITERATURE REVIEW</td>
<td>9</td>
</tr>
<tr>
<td>2.1 Comparing Aberration Detection Methods with Simulated Data</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Syndromic Surveillance: STL for modeling, visualizing, and monitoring disease counts</td>
<td>11</td>
</tr>
<tr>
<td>3 LOCALLY WEIGHTED SCATTERPLOT SMOOTHER (LOESS)</td>
<td>14</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>14</td>
</tr>
<tr>
<td>3.2 Model specification</td>
<td>15</td>
</tr>
<tr>
<td>3.3 Interpolative data blending procedures</td>
<td>18</td>
</tr>
<tr>
<td>3.4 Characterization of the distributional properties of the variance</td>
<td>21</td>
</tr>
<tr>
<td>3.5 Robust LOESS estimation</td>
<td>23</td>
</tr>
<tr>
<td>3.6 Accounting for seasonality in epidemiological data</td>
<td>24</td>
</tr>
</tbody>
</table>
4 DATA ANALYSIS (DATA SET #1).............................................................................. 27
4.1 Making the case for simulated data................................................................. 27
    4.11 Description of simulated data sets.............................................................. 27
    4.12 Superimposition of outbreaks..................................................................... 28
4.2 Corrective data transformations..................................................................... 29
    4.21 Establishing a starting point........................................................................ 31
    4.22 Model selection criteria for optimal pooling candidates.......................... 33
    4.23 Residual diagnostics.................................................................................. 36
4.3 Dynamic pooling methodology....................................................................... 38
4.4 Post-pooling adjustments............................................................................... 40
4.5 Creating an alarm threshold........................................................................... 41
4.6 Initial results................................................................................................... 43
    4.61 Initial results excluding one day spikes..................................................... 43
4.7 Simulated out-of-sample performance............................................................ 44
4.8 Concluding results for data series #1............................................................. 46

5 DATA SET #2 (EXTREME SEASONALITY).......................................................... 47
5.1 Corrective data transformations..................................................................... 47
5.2 Note on seasonal adjustment......................................................................... 47
5.3 Selection criteria for optimal model pooling candidates.............................. 50
5.4 Residual diagnostics...................................................................................... 53
5.5 Initial results for data series #2...................................................................... 56
    5.51 Results excluding one day outbreaks......................................................... 57
5.6 Simulated out-of-sample performance............................................................ 58

6 CONCLUSION.................................................................................................... 60
Bibliography.......................................................................................................... 63
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.6a Efficiency frontier data table</td>
<td>34</td>
</tr>
<tr>
<td>4.7a LOESS model performance comparison</td>
<td>35</td>
</tr>
<tr>
<td>4.11a LOESS vs pooled model performance comparison</td>
<td>43</td>
</tr>
<tr>
<td>4.11b LOESS vs pooled model performance comparisons sans 1-day outbreaks</td>
<td>43</td>
</tr>
<tr>
<td>4.13a Simulated out-of-sample performance comparison</td>
<td>45</td>
</tr>
<tr>
<td>4.13b Simulated out-of-sample performance comparison sans 1-day outbreaks</td>
<td>46</td>
</tr>
<tr>
<td>5.3a Efficiency frontier data table sans 1-day outbreaks</td>
<td>50</td>
</tr>
<tr>
<td>5.9 Pool model vs LOESS performance comparison</td>
<td>56</td>
</tr>
<tr>
<td>5.10 Pooled model vs LOESS performance comparison sans 1-day outbreaks</td>
<td>57</td>
</tr>
<tr>
<td>5.12 Simulated out-of-sample model performance comparison</td>
<td>58</td>
</tr>
</tbody>
</table>


### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Cropped picture of model comparison results from Hutwagner et al.(2005)</td>
<td>10</td>
</tr>
<tr>
<td>2.2 Model performance control chart taken from Hafen et al. (2009)</td>
<td>12</td>
</tr>
<tr>
<td>3.1 Graphical display of kd-tree orthogonalization process</td>
<td>19</td>
</tr>
<tr>
<td>4.1 Dependent variable square root transformation</td>
<td>29</td>
</tr>
<tr>
<td>4.2 Transformed dependent variable de-trended</td>
<td>30</td>
</tr>
<tr>
<td>4.3 Transformed dependent variable de-cycled</td>
<td>30</td>
</tr>
<tr>
<td>4.4 AICc optimized smoothing parameter vs dependent variable</td>
<td>31</td>
</tr>
<tr>
<td>4.5 AICc optimized smoothing parameter model residual diagnostics</td>
<td>33</td>
</tr>
<tr>
<td>4.6b Efficiency frontier graph</td>
<td>34</td>
</tr>
<tr>
<td>4.7b .02 LOESS smoothing parameter vs dependent variable</td>
<td>35</td>
</tr>
<tr>
<td>4.7c .05 LOESS smoothing parameter vs dependent variable</td>
<td>36</td>
</tr>
<tr>
<td>4.8a .02 LOESS smoothing parameter residual diagnostics</td>
<td>36</td>
</tr>
<tr>
<td>4.8b .05 LOESS smoothing parameter residual diagnostics</td>
<td>37</td>
</tr>
<tr>
<td>4.9a Pooled prediction series vs dependent variable</td>
<td>39</td>
</tr>
<tr>
<td>4.9b Histogram for pooled prediction series residuals</td>
<td>40</td>
</tr>
<tr>
<td>4.10 Accuracy vs false positives graph</td>
<td>42</td>
</tr>
<tr>
<td>4.12 Simulated out-of-sample series predictions</td>
<td>45</td>
</tr>
<tr>
<td>5.1 Square root of transformed, decycled dependent variable with seasonality</td>
<td>48</td>
</tr>
<tr>
<td>5.2 Seasonally adjusted data series</td>
<td>49</td>
</tr>
<tr>
<td>5.3b Efficiency frontier graph</td>
<td>51</td>
</tr>
<tr>
<td>5.4 .02 &amp; .05 LOESS model predictions vs dependent variable observations</td>
<td>52</td>
</tr>
<tr>
<td>5.5a .02 LOESS model residual diagnostics</td>
<td>53</td>
</tr>
<tr>
<td>5.5b .05 LOESS model residual diagnostics</td>
<td>53</td>
</tr>
<tr>
<td>5.6 Pooled prediction series vs dependent variable</td>
<td>54</td>
</tr>
<tr>
<td>5.7 Histogram of pooled model residuals</td>
<td>55</td>
</tr>
</tbody>
</table>
5.8 Model accuracy and false-positive rates................................................................. 56
5.11 Simulated out-of-sample series predictions......................................................... 58
CHAPTER 1

INTRODUCTION TO SYNDROMIC SURVEILLANCE

1.1 Introduction

Improving syndromic surveillance methods for public and animal health systems has become a hot button topic ever since the series of biological terrorist attacks in 2001 utilizing anthrax were carried out through U.S. postal delivery services. At first, no one knew how widespread the attack was, and so the problem facing public health officials was identifying new areas where outbreaks were occurring as quickly as possible, so that the outbreaks could be contained and treatment could be administered to those infected. The scientific community found itself woefully ill-prepared to carry out such a task; since that time syndromic surveillance research has intensified and diversified so that the next crisis will not catch us unawares. Syndromic surveillance is defined generally as the collection and statistical analysis of data which are believed to be leading indicators for the presence of pernicious activities developing within or acting upon a system of interest. By this definition, relevant data could range from the number of pills of Nyquil sold in a given region if trying to predict a flu epidemic, to the number of patients complaining of respiratory and flu-like symptoms in emergency rooms, which are believed to be the most likely early symptoms victims of a biological terrorist attack might first experience (Franz, 1997). The greater challenge facing practitioners wishing to employ syndromic surveillance systems, however, is not the identification of data which are relevant sentinel predictors of an outcome of interest, but defining the detection algorithms which determine whether or not an observed data value is aberrant to the point that it is flagged for investigation. The speed and accuracy with which an outlier caused due to an outbreak is
flagged is measured by the detections system’s sensitivity and mean time to detection, while the number of false-positives, or the number of false-alarms, is measured by its specificity.

1.2 Why syndromic surveillance?

Much of the interest and attention devoted to developing and improving syndromic surveillance systems is derived from two major trends within human civilization, whose prominence and relevance are expected to continue increasing for the foreseeable future. The first factor is the paradigm shift in daily life caused by the unprecedented dynamism of modern information technology; the world is increasingly instrumented, with computers and data storage devices improving in speed, storage capacity, and battery life while at the same time becoming smaller and more lightweight. Additionally, these devices are also increasingly networked and capable of communicating with each other at increasingly faster speeds and larger bandwidths. This means mountains upon mountains of data, figuratively speaking, are created and stored every day. Google and AT&T combine to create or transfer an average of 43 petabytes of data per day alone\(^1\); put in terms of a more commonly used metric, these two companies alone account for more than 43,000,000 gigabytes of data per day. This means complete data sets from farms, hospitals, field workers, veterinarians, and virtually any other source imaginable is available almost instantaneously and updated in near real-time. The real challenge therefore lies in using this abundance of data intelligently in such a way that the underlying processes generating them can be identified, and using the benefits of insights gained from such analysis, inform decision makers when abnormal activity that warrants closer investigation appears in the system.

\(^1\) http://en.wikipedia.org/wiki/Petabyte
An example of a database which would be ideal for the application of syndromic surveillance is the Emergency Management Response System (EMRS), developed and maintained by the U.S. Department of Agriculture (USDA), Animal Plant and Health Inspection Service (APHIS), Veterinary Services, and the Centers for Epidemiological and Animal Health (CEAH). The EMRS is a nationally integrated database meant to provide the administrative capabilities necessary to track when and where a disease was identified, what response action was taken, when and how the issue was followed up, costs to the agency both in executing the response and, if applicable, compensating the farmer. It allows for real-time case reporting by field agents, which can then be tracked or indexed by disease type, animal type, geographic location, time period, or any other number of potential groupings of interest. The EMRS database can also be accessed by field workers using laptops or smart cell phone technologies to maximize the speed of information transfer. This database is at the heart of the information technology machinery the U.S. utilizes to manage the accounting, administration, and execution of disease surveillance, outbreak tracking, response strategy, and post-facto performance evaluation. The system makes available data regarding historical outbreaks, administrative costs, and response and control costs that are both broad and deep. It is crucially important that EMRS reports both the outcome of actions taken in response to a potential threat, as well as the costs associated with that response. This makes it possible to estimate the average cost per investigation and the marginal cost per additional investigation. In order for a monitoring agency to exercise control of its budget, it is important to know how many false-positives are financially feasible for the agency to investigate, on average, over a period of time. As will be shown later, the ability to identify costs in order to create a budget constraint as a function of the maximum acceptable number of values incorrectly flagged as outbreaks (false-positives) is very important.
The other major impetus for the accelerated development and deployment of surveillance systems is the threat of bioterrorism and naturally occurring human and animal health epidemics. Some of the most well known and damaging animal-based epidemics include Avian Influenza (H7N2-2004, H5N1-1997), SARS, Foot and Mouth Disease, and Bovine Spongiform Encephalopathy. The common causal trait underlying all of these outbreaks is that of poor hygiene and lack of adherence to basic guidelines of common sense cleanliness, although the reasons for non-compliance vary.

This is an increasingly relevant topic, as three quarters of all human disease which are emerging or re-emerging as of the beginning of the twenty first century are caused by pathogens originating from animals or products of animal origin (Willingham, 2005). The very things which would seem to have heralded man’s mastery over nature— the advent of megacities, global travel, modern medicine, etc., have all ultimately contributed to the growing threat of epizootic and zoonotic epidemics. The distinction between the two is that zoonotic diseases are pathogens that originally began in animal reservoirs but which through environmental mutation become capable of infecting humans, while epizootics are widespread outbreaks of disease in animal populations that are not capable of infecting humans. The relationship between the two is that most zoonotics start as epizootics, and then mutate to become transmissible to humans. The more animals infected in an epizootic, the greater the likelihood that somewhere a perfect storm of circumstances will lead to the epizootic strain mutating into a zoonotic one; it only needs to happen in once.

1.3 Historical Instances of Epizootics

The index case for the outbreak of Food and Mouth disease in the United Kingdom was attributed to feeding swine contaminated waste food. Feeding industrial garbage to commercial
livestock is a long standing practice in the meat industry, and was adopted, presumably, in response to increasing pressure to produce more food at lower cost. Waste food can be procured from restaurants, cruise ships, military bases, etc., and must be processed, typically through a combination of mechanical grinding and “cooking” through exposure to high levels of heat, in order to qualify as being fit for animal consumption. These processing measures are widely regulated by the relevant organs of the state in countries with mature economies, and are supposed to theoretically eliminate all harmful bacteria and pathogens. Upon investigation it was discovered that the originating farm had been placing unprocessed waste food in feed barrels labeled as processed waste-feed only.

The onset of Bovine Spongiform Encephalopathy (BSE), also known as “Mad Cow Disease”, in the United Kingdom was also attributed to negligent farm feeding practices. Cattle, which are naturally free-grazing herbivores, are widely fed diets consisting mainly of corn and waste-food because the commercialization of the cattle industry has eliminated the practice of grazing for industrial meat-packing operations. In the outbreak of BSE in the UK, cattle were fed the meat and bone meal (MBM) of other cattle that had previously been slaughtered but for one reason or another been considered unfit for human consumption. Such reasons could include disease, lameness, excessive morbidity leading to the loss of the ability to stand, etc. The recycling of meat and bone meal feed back in to the herd of the same type from which it came was also a long standing and widespread practice of commercial agriculture. Such practices have since been banned by most developed nations, however in developing parts the world where regulation is more lax, such as China and India, such food control safeguards do not exist. Due to the horrifying effects of the illness in humans in the form of new variant Creutzfeldt-Jakob disease and the near certain death of those who contract it, countries witnessing outbreaks of BSE have noticed significant declines in domestic and foreign demand for meat.
Outbreaks in poultry, specifically SARS and Avian Flu, are attributed to a variety of casual factors, but the number one contributor is indisputably chronically poor sanitation practices on the part of small and medium sized farms. Poor sanitation combined with a propensity to keep all live animal stocks (ducks, chickens, roosters, rabbits, etc.) in a common housing, as well as frequent close contact with human handlers, leads to a situation which is ripe for the emergence of zoonoses. The outbreak of SARS in 2003 was attributed to the province of Guangdong in southern China, where the agricultural and industrial sides of modern China are often no more than a few hundred meters of open air apart from each other. The potential for livestock holding pens to act as host reservoirs for new zoonoses and epizootics is exacerbated by this proximity to industrial manufacturing plants, which in China and other quickly industrializing nations are well known for their lack of adherence to safety regulations if doing so increases costs or slows production.

All of this indicates that in order for our first line of defense against animal disease should be preemptive rather than reactionary; animal populations should be guarded against exposure to epizootics through syndromic surveillance systems which quickly identify outbreaks so that they cannot spread. In doing, the human population will indirectly be better protected from the threat of a virulent new zoonotic disease developing out of the Petri dish of an epizootic outbreak.

1.4 A Dangerous Melting Pot

From 1950 to 2009, the global population more than doubled from 2.9 billion to 6.75 billion people, and is expected to top 9 billion by the year 2050\(^2\). This has led to vast increases in

\(^2\) http://www.census.gov/ipc/www/idb/worldpopgraph.php
demand for animal based protein and associated products, which in turn has forced radical changes in animal husbandry practices, necessitated by the need to produce larger numbers of animals in smaller amounts of space. In 2007, more than 21 billion animals (more than three times the global population of humans during that same time period) were produced, and demand for animal protein is expected to grow 50% by 2020 (Lemon, 2007). The three main stocks which fuel this growth, namely swine, poultry, and cattle, have unsurprisingly been the three animal reservoirs in which the majority of zoonotic diseases have developed during the past sixty years. The majority of this human population growth is taking place in undeveloped countries, where illiteracy and abject poverty are highest, and standards of hygiene are lowest. Much of this population increase will be absorbed by megacities, defined as metropolitan areas characterized by extremely high population densities and habitation counts in excess of ten million people. By 2030, sixty percent of the world’s population is expected to live in cities, and as early as 2015, there will be 22 megacities, 17 of which will be located in developing countries.

Some 800 million poor worldwide depend on livestock for their daily sustenance and survival; this number represents roughly 8.5% of the current global population, which although sizeable in and of itself, is even more staggering when one considers the fact that rather than being evenly distributed around the world, these 800 million people are highly concentrated in specific regions, accounting for 70% of the population in Africa and 90% of the population in Asia.

All of these factors highlight the need for the public health institutions that safeguard the health of both humans and animals to remain vigilant in safeguarding against the emergence of new diseases. To accomplish this, organizations must be equipped with the tools necessary to deal with the challenges they face. Towards aiding in the success of such endeavors, the data used for this study are simulated counts of clinical signs believed to be
leading indicators for the presence of a pernicious disease in an animal population. In the context of epizootics and animal health, these are the types of symptoms which are expected to be seen as early warning signs of a greater outbreak occurring. This could range from data on animal morbidity reports to the count of field sightings for a symptom known to be generally indicative of a particular pathogen. The problem is that sickness and disease are a natural part of life, and so a syndromic surveillance system, in this context, must be able to identify when a given system is departing from the reasonable bounds of its baseline, signaling an outbreak may be occurring that warrants further investigation. To accomplish this, syndromic surveillance systems must rely upon aberration detection algorithms to determine whether an observation for a given period is abnormal enough to warrant being flagged for investigation.
CHAPTER 2
LITERATURE REVIEW

2.1 Comparing Aberration Detection Methods with Simulated Data

The idea for this thesis began with a paper called *Comparing Aberration Detection Methods with Simulated Data* (Hutwagner et al., 2005), written by a team of Centers for Disease Control and Prevention (CDC) statisticians. This paper compares the sensitivity, mean time to detection, and specificity of five different models, three of which are very similar: C1-Mild, C2-Medium, and C3 Ultra, which are based on positive one-sided cumulative sum control chart (CUSUM) calculations and are the standard detection algorithms used by the CDC. Models C1 and C2 have aberration detection flag thresholds at the mean +3 standard deviations. These methods are all essentially 7 day moving averages, some of which are lagged behind the current date being estimated. The test statistic derived takes the form:

\[
S_t = \max(0, y_t - \frac{u_t + \sigma_t}{\sigma_t}),
\]

where \(u_t\) and \(\sigma_t\) are a 7-day baseline mean and standard deviation. The C2-MILD and C3-Ultra models are lagged 2 days behind the current time period, with C2-Medium sharing the same threshold as C1-Mild, which is not lagged at all. The C3-Ultra alarm threshold is 1.5 standard deviations away from its mean. The models were compared on 50 different possible combinations of baselines with varying mean, variance, seasonality, and trend. Ten thousand iterations of a six year period from Jan 1\textsuperscript{st}, 1994 to Dec 31\textsuperscript{st}, 1999 corrected for holidays and leap-years, was then conducted for each of the 50 baselines. The models’ performance was then compared. The exact performance metrics should be clarified to provide a better understanding of the figures below.
Sensitivity = \frac{\sum b}{m} \text{ where } b \text{ is the binary variable } [0,1] \text{ and } m \text{ is the number of total outbreaks which occurred in the data set. This is the model’s ability to identify an outbreak. If a model correctly flags an outbreak in at least one period, the entire outbreak is considered a 1, or a 0 if the model never flagged an outbreak at any period.}

Mean Days Until Detection measures the speed with which a model identifies an outbreak correctly. The data are presumed to become available only after a full day has passed, therefore if an outbreak begins on day 1 and is identified on day 4, the time to detection is 4 because the data wouldn’t have been available until the final day’s count had been reported in the evening.

Specificity = (1 - \frac{\sum f}{n}) \text{ where } f \text{ is the number of times the model flagged an observation as an aberration when no disease outbreak was present, and } n \text{ is the sample size (6 years=2191 days, 60,000 years=21,910,000 days). } \frac{\sum f}{n} \text{ itself is the false-positive rate.}

Figure 2.1: Cropped picture of model comparison results from Hutwagner et al., 2005³

Original table available in full at http://www.cdc.gov/ncidod/EID/vol11no02/04-0587_app1.pdf
The basic trend here is that the less sensitive models, e.g., C1-MILD, consistently have very high specificities and the lowest mean times to detection, while the more sensitive models, e.g., C3-ULTRA, have higher sensitivities but sacrifice in the specificity and mean time to detection metrics. The mean time to detection must be considered relative to the sensitivity of the model however; although C1-Mild may have the lowest mean time to detection, the number of outbreaks it’s actually detecting may be significantly lower than C3-ULTRA. Overall, the models seem to show poorer performance across the board when applied to series without systematic variation, i.e., no trend and low/no seasonality, as well larger values for the means and standard deviations of the generating parameters. The models generally perform best when significant systemic variation is present and/or the mean and variance are small. The comparative statics for the model behavior are as expected; as sensitivity goes up, specificity suffers, and vice versa. Looking at the figure above, the loss in specificity would probably be outweighed by the sensitivity gains exhibited by the C3-ULTRA method, but even in that case, in some instances where there is no trend or seasonality present and a medium to large mean and standard deviation, as highlighted in the figure above, even the C3-ULTRA misses 25-30% of outbreaks, which may be an unacceptably high number depending on the gravitas of the disease being monitored.

2.2 Syndromic Surveillance: STL for modeling, visualizing, and monitoring disease counts

The research article *Syndromic Surveillance: STL for modeling, visualizing, and monitoring disease counts* (Hafen et al., 2009) compares the model performance of the C1, C2, and C3 methods with a general linear model (GLM) and seasonal-trend decomposition procedure with loess (STL). The general linear model is essentially a fixed effects model with interactive dummy variables for days of the week (Monday-Sunday), months of the year (Jan-
Dec), and a linear time component to account for long term trend. The STL method is a procedure which decomposes the systematic sources of variation in the dependent time series variable, specifically trend, cycle, and seasonality, and then estimates a LOESS (also known as LOWESS: LOcally WEighted Scatterplot Smoother) model, which is a form of local regression to be discussed further, to flag abnormal values for investigation.

As can be seen in figure 2.2, although GLM comes close in some cases, STL and STL(90), the latter of which only uses data from the previous 90 days for estimation, show better overall detection capabilities than any of the other models. C1, C2, and C3 do not even come close to being competitive. The question most analysts would be left with after viewing this table would probably be whether to move forward using the full-history STL or the STL 90-day as the chosen aberration detection algorithm. This is a difficult decision because no model is clearly dominant; out of the nine combinations of baselines and magnitudes, the STL method displays higher sensitivity in 5 instances, the STL(90) sensitivity superior in 3 instances, and in one case they are identically tied. Neither the STL nor STL (90) models display clear dominance in the mean time to detection metric; observed model specificities were not reported by the authors. The results

---

4 STL and GLM model performance metrics calculated with a 97% confidence interval. Observations outside the confidence interval are flagged as outbreaks.
presented in each of these papers leave the analyst tasked with choosing a detection algorithm with a certain amount of ambiguity—which model should be chosen, when each method has its strengths and weaknesses, and in the case of the STL methods, none are clearly dominant in all scenarios? Seeking the answer to this question is what ultimately led to the development of the research question, “How can these individual models be combined so that the pooled ‘child’ model retains the strengths of its parents but not their weaknesses?” The LOESS modeling methodology, having shown superiority to the C1, C2, and C3, was selected as the basis from which the individual models used in the pooling process would be created. Before inquiry in to the research question can begin in earnest however, a comprehensive description of LOESS, a non-parametric local regression methodology, should be explained.
CHAPTER 3

LOCALLY WEIGHTED SCATTERPLOT SMOOTHER (LOESS)

3.1 Introduction

The modeling methodology used for creating the aberration detection algorithms employed throughout this analysis is a relatively new approach to fitting regression surfaces; local regression analysis. The early beginnings of this method (Macauley, 1931) suggested that in order to provide a smoothed estimate for a regression surface at time t, a polynomial be fitted over an interval centered at time t, from which an estimate is calculated by taking the solution values of the fitted polynomial evaluated at that point. In the special case that a data series is uniformly distributed and the points being estimated are not too close to the beginning or end of the data set, the set of smoothed prediction values are calculated identically to a moving average with fixed coefficients for the data points included in the interval. This is conceptually very similar to what a local regression does.

Local regressions as they are used today were developed by William S. Cleveland in 1979, an AT&T Bell Laboratories econometrician, when he suggested estimating the regression surface at a point x by fitting a linear or quadratic polynomial function of the independent variables, centered and evaluated at x. The size of the neighborhood of data points used to estimate the polynomial function is controlled by adjusting the smoothing parameter, which determines the exact locality of the regression. The effect that the smoothing parameter has on the smoothness of the estimated series is analogous to changing the bandwidth of a moving average; a smaller local neighborhood of data points results in a more jagged series of estimates that interpolate the observed values of the data, while a larger neighborhood is much more smooth and less prone being influenced by outliers. The impetus that drove Cleveland to develop local regressions was the desire to create a new regression paradigm that allowed for a
larger class of possible regression surfaces than ordinary least squares, but that could be estimated without losing the crucial statistical properties that common inferential methods rely upon. Not all regression surfaces can be well approximated by the low-order polynomials normally associated with least squares, yet local regression has shown it can fit almost any series without much distortion. The other major benefit of local regressions is that they are non-parametric; a LOESS model can be constructed using as little as the observations of a dependent variable and a linear time component. This means specifying the explanatory variables, and worrying about things like multicollinearity, endogeneity, specification bias etc., are not problems which LOESS has to contend with. Although inherently non-parametric, if desired multiple explanatory variables and their respective neighborhoods can be utilized in the estimation process.

3.2 Model specification

A simple local regression model would look something like this:

Let $Y_i$ be the $i$th observation of the dependent variable, and $x_i$ be the $i$th observation of the independent variable. The specification of the local regression model would be:

$$Y_i = g(x_i) + u_i,$$

where

Three assumptions are made: (1) $\hat{g}(x)$ is an unbiased estimate of $g(x)$, (2) $g(x)$ is a smooth, continuous function of $x$, and (3) that the residuals are normal with mean zero and constant variance, implying the use of 4 of the 7 classical linear regression model assumptions, $E[u|x_i] = 0$, $\text{Var}(u|x_i) = \sigma^2$, and $u \sim \text{iid normal}$. The notation $\hat{g}(x_i)$ is used because the values of $x_i$ at any
point are inserted into a continuously defined function throughout the neighborhood that maps its value, which is described in greater detail shortly. The important thing to recognize is that it is not the actual values of the \( x_i \)'s which are used to directly predict \( \hat{y}_i \), but the values of the function \( \hat{g} \), which are determined by \( x_i \), combined with observations of the dependent variable.

The dependent variable \( Y_i \) is a linear function \( g \) of the explanatory variable \( x_i \) resulting in a linear model such that:

\[
\hat{g}(x) = \sum_{i=1}^{n} l_i(x) y_i
\]

The form of the local regression model allows for the retention of the crucial classical regression model assumptions about unbiasedness and the structure of the disturbance, allowing for the estimation of variance and the use of confidence intervals and hypothesis tests.

The \( l_i \) in the equation above refers to the unique weighted value of the \( x_i \)th observation at that point. The weight assigned to the \( x_i \)th observation of the independent variable results from coefficient estimated by least squares after the observation has been transformed by a weighting and smoothing procedure. The initial weighting is done through applying a tri-cube loss function which exponentially decreases the weighting assigned to values as they grow more distant relative to the value being estimated, where the neighborhood is centered.

In the univariate case, the smoothing function is defined as follows—Let:

\( d(x) \) be the distance of \( x \) to the \( q \)th nearest \( x_i \), and let

\[
w_i(x) = W\left(\frac{|x_i - x|}{d(x)}\right)
\]

be the weight of the \( x_i \)th observation of the independent variable where the weighting function \( W \) takes the form of the tri-cube function: \( W(\mu) = (1-\mu^3)^3 \) for \( 0 \leq \mu \leq 1 \).
1. The form of the weighting function $W$ is one of the choice parameters available to be changed freely by the analyst; for a discussion on asymmetric weighting kernels, see Bianconcini, 2007.

In the multivariate case, there are $k$ independent variables so that each $x_i$ is part of a $k$-dimensional vector of values. If $k$ is assumed to be a distance function in this space, then the new smoothing function becomes:

$$w_i(x) = W \left( \frac{k(x_i - x)}{d(x)} \right)$$

The distance function $k$ in this case is calculated by scaling the independent variables by individually dividing them by their interquartile range and using the Euclidean distance of the scaled variables.

Two further decisions must be made before estimation can proceed. The first is to choose the value of the smoothing parameter, $\beta$, which will define the size of the local neighborhood of data points used in local regression. This number must be between 0 and 1, and represents the proportion of data used in each local regression neighborhood relative to all data in the sample space. As mentioned previously, the smoothness of the function $\hat{g}(x)$ depends directly on size of the local neighborhood analogously to how the smoothness of a moving average is determined by its bandwidth. In general, as $\beta$ increases, the bias of $\hat{g}(x)$ tends to grow and the variance tends to shrink. Methods for selecting the optimal value for the smoothing parameter are explained later.

The final decision is to choose the degree of the polynomial functions to be fit over the local regression space, which can be either linear or quadratic. Locally linear polynomial fitting works best on dependent variables which are mostly smooth functions without sufficient curvature to produce multiple local minima and maxima. Fitting linear polynomials to data
series that display considerable curvature results in substantially biased predictions even for small values of $\beta$. Locally quadratic polynomial fitting is the superior choice for such data, although there is a trade off to be made in terms of the degrees of freedom available for estimation because quadratic fitting includes the squared and cross product transformations of the independent variables, leading to $2p + p(p-1)/2$ fitting variables for $p$ independent variables. Moving from linear to quadratic local fitting produces a similar effect to adding additional explanatory variables to a classical regression model; bias shrinks while variance grows.

### 3.3 Interpolative data blending procedures

The points in the neighborhood are used in either a linear or cubic interpolative blending process that creates a continuously defined function throughout the neighborhood. To ease the computational burden of estimation, especially for large data sets, the blending function is not estimated using all points in the local neighborhood. Investigating the behavior of the smoothed surface utilizing all data in the local neighborhood revealed that the surface changes slowly, so it is possible to move forward without using every data point in the local neighborhood for the blending process while still maintaining a high degree of accuracy and being much more computationally efficient. To achieve this, a data segmentation process takes place that chooses which data points in the neighborhood will be used in the blending process.

This is accomplished by partitioning the space of the independent variables using a k-d tree method (Friedman, Bentley, and Finkel, 1977), where only the points on the vertices of the rectangles shown below are evaluated. A k-d tree is a hierarchical data partitioning process that consists of successive orthogonal cuts to the coordinate axes in $k$ dimensions, where $k$ is the vector of independent variables, as shown in Figure 3.1 below. This process is designed to
compensate for the fact that there may be local neighborhoods whose distributions are not normally distributed.

Imagine that you know a function $g(\mu, \nu)$ on the boundary of a rectangle where $0 \leq \mu \leq 1$, $0 \leq \nu \leq 1$. Let:

$$s(u, v) = \sum_{\xi=0.1} \phi_{\xi}(u)g(\xi, v) + \sum_{\nu=0.1} \phi_{\nu}(u, \nu) - \sum_{\xi=0.1} \sum_{\nu=0.1} \phi_{\xi}(u)\phi_{\nu}(v)g(\xi, \nu)$$

be the linear interpolative blending function.

The first term on the right hand side of the equation calculates values over the range $[0,1]$ of the horizontal axes, the second term calculates values over the range $[0,1]$ of the vertical axes, and the final term subtracts a compensating bilinear interpolant from the four corners. All possible combinations of values for $u$, $v$, and $\xi$ are calculated, forming a one-dimensional surface along the intersection of the horizontal and vertical axes of the rectangles identified by the kd-tree.
orthogonalization process. This smoothes the values in the local neighborhood into a continuously defined function, providing the added convenience that missing data values do not adversely impact model estimation when using this blending method. The second interpolation method involves solving a unique cubic polynomial equation whose solution values, when the first derivatives of the cubic function are set to 0, match those of the fitted local polynomials evaluated at the two endpoints of the kd tree rectangles. The unique cubic functions whose solution values match the vertices of each kd-tree are then selected to estimate a continuous surface. The cubic interpolation method provides slightly better accuracy than linear interpolation in most cases, and with modern computing technology making the increased computational burden negligible, is generally the preferred method.

Because $\hat{g}(x)$ is a linear function of the independent variables, it’s possible to characterize its variance with familiar equations.

$$\hat{g}(x) = \sum_{i=1}^{n} l_i(x)y_i$$

$$\sigma^2(x) = \sigma^2 \sum_{i=1}^{n} l_i^2(x)$$

Let $\hat{y}$ be the vector of fitted values, $y$ the vector of dependent variable observations, and $\hat{\epsilon}$ the vector of residuals. Then:

$$\hat{y} = Ly, \text{ and}$$

$$\hat{\epsilon} = (I - L)y$$

Where the $(i, j)$th element of $L$ is $l_j(x_i)$. The distributional properties of the residuals associated with the classical linear regression model assumptions are maintained. $L$ is the matrix of weights which minimize the residual sums of squares over the surfaces estimated by the
weighting, smoothing, and interpolation processes mentioned previously, and is analogous to the familiar beta vector \((\beta)\) of estimated partial effects in an OLS model. \(L\) shares the property with its OLS counterpart that if \(z\) is a vector in this space, then \(Lz=z\), but differs in that it is not symmetric or idempotent.

### 3.4 Characterization of the distributional properties of the variance

The variance-covariance estimates for the model are defined as:

\[
\sigma^2 LL' \quad \text{and} \quad \sigma^2 (I - L)(I - L)'
\]

To create confidence intervals and conduct hypothesis testing, let:

\[\varphi_k = \text{Trace}[(I - L)(I - L)']^k \quad \text{for } k = 1, 2\]

Then, since

\[E[\sum_{i=1}^{n} \delta_i^2] = \sigma^2 \varphi_1, \quad \sigma^2 \text{ can be estimated by } \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \delta_i^2}{\varphi_1}\]

Confidence intervals can be estimated using the result that a constant times a sum of squares is approximately distributed as a \(\chi^2\) \((\text{Kendall and Stuart, 1977})\), where the constant and the degrees of freedom are chosen so that the first two moments of the constant, in this case \(\varphi_k\) for \(k = 1, 2\), times the sum of squares match those approximating the \(\chi^2\) distribution. Then:

\[
\frac{\varphi_1^2 \hat{\sigma}^2}{\varphi_2^2 \sigma^2}
\]

is approximated by a \(\chi^2\) distribution with \(\frac{\varphi_1^2}{\varphi_2}\) degrees of freedom.

The test statistic:

\[
\frac{\hat{\beta}(x) - \beta(x)}{\sigma^2}
\]
is then approximated by a t-distribution with $\frac{\varphi_1^2}{\varphi_2}$ degrees of freedom.

There are three possible selection criterion methods which can be used for choosing the smoothing parameter $\beta$. The first two take the form of modified Akaike Information Criterion:

$$\text{AICc}_1 = n \log(\hat{\sigma}^2) + n \left( \frac{\varphi_1}{\varphi_2(n+\varphi_1)} \right) \frac{\varphi_1}{\varphi_2-2}$$

$$\text{AICc} = \log(\hat{\sigma}^2) + 1 + \left( \frac{2(\text{Trace}(L) + 1)}{n - \text{Trace}(L) - 2} \right)$$

The third is a Generalized Cross Validation (GCV) criterion:

$$\text{GCV} = \frac{n\hat{\sigma}^2}{(n - \text{Trace}(L))^2}$$

All three of these methods seek to optimize the balance between parameter parsimony and the explanatory power of the model. For most purposes, the value of the smoothing parameter $\beta$ which minimizes one of these functions is typically a good starting point for model estimation. Another method for choosing the smoothing parameter involves creating an M-plot (Cleveland, 1988) which graphically shows the trade-off between bias and variance, enabling the analyst to make an informed decision about the value they select for the smoothing parameter based on whether they place a premium on unbiasedness or minimum variance.

In summation, to estimate a local regression model, the analyst must choose the values for $W$, the neighborhood weighting function, the smoothing parameter $\beta$, and the degree of the polynomial (either linear or quadratic) functions used to estimate the regression surface. When this process commences, the values in the local neighborhood being estimated are modified by
the tri-cube weight function, and the data points in the neighborhood used to interpolate a continuous surface are selected through the kd-tree segmentation process. These weighted, smoothed values are either left alone in the case of linear polynomial fitting, or the squared and cross products of the data are added in the quadratic case, and the familiar least squares estimation process is performed.

3.5 Robust LOESS estimation

It should be noted that there exists a robustification procedure, which results in “Robust LOESS” model estimation, which modifies the weights originally assigned through application of the tri-cube loss function in an effort to reduce the adverse effects of outliers on the fit of the local regression neighborhood. This is accomplished by inserting the observed residuals, \( \hat{\epsilon}_i \) for \( i=1,\ldots,n \), in to the following formula:

\[
\rho_i = B\left(\frac{\hat{\epsilon}_i}{s}\right)
\]

Where \( B \) is the bi-square function defined as: \( \beta(u) = (1 - |u|^2)^2 \) for \( u < 1 \), and \( s \) is the median value of the previously estimated model residuals. The form of this function means that estimates with larger residuals are given smaller weights in future estimation. The model is then re-estimated after multiplying the weights originally assigned by the tri-cube loss function with the values \( \rho_i \) derived from the robust weighting function to form a new weight, \( \hat{w}_i(x) = \rho_i w_i(x) \). This process can be done recursively as many times as desired, although in practice it has been generally observed that 1 or 2 iterations are sufficient for the robust weights to stabilize.
3.6 Accounting for seasonality in epidemiological data

There also exists a body of work nearly as extensive as that of detection algorithms regarding the optimal way to deal with seasonality in the context of identifying disease outbreaks. When dealing with normal economic variables such as quantities demanded, prices, etc., extracting the systemic variation of seasonality is a fairly straightforward process because the seasonal component typically manifests itself with nearly identical periodicity according to calendar year. One can always expect retail sales to be significantly higher during the major holiday purchasing seasons of November and December than the summer months of June, July, and August, for example. Unfortunately, forces of nature such as germs and disease do not provide such predictable conformity. While it is well known for instance that cases of the flu increase during the months of November, December, January, and February in climates where it becomes much colder, the exact peaks and troughs of the outbreaks show considerable variability from year to year. In the case of seasonal flu, instances have been recorded in which the pattern of outbreaks is actually bi-modal.

One method for an epidemiological seasonal adjustment mechanism would involve the application of a non-hierarchical kernel smoothing process such as the one below:

\[
c(d) \sim \text{Poisson} \left( \sum_{i=1}^{7} c_i l_i(d) \right) + \left[ c_8 + c_9 d \right] + \left[ c_{10} \cos \left( \frac{2\pi d}{365.25} \right) + c_{11} \sin \left( \frac{2\pi d}{365.25} \right) \right]
\]

Where \( \sum_{i=1}^{7} c_i l_i(d) \) extracts the day-of-the-week effect where \( c_i \) represents the weight given on day \( i \) for \( i = 1, \ldots, 7 \), and which must sum to 7, \( l_i \) is a binary variable which takes the value 1 if \( d_i = d \), and 0 otherwise. \( c_8 + c_9 d \) represents the intercept and long-term trend effects respectively, and \( c_{10} \cos \left( \frac{2\pi d}{365.25} \right) + c_{11} \sin \left( \frac{2\pi d}{365.25} \right) \) represent the estimated seasonal effects. A “kernel” smoothing process is simply a series of estimated smoothing weights that are
mechanically applied to a data set after estimation with no further learning. A moving average could be considered a special case of a kernel smoother in that it applies a uniform weight which sums to 1 to all of the values contained in its bandwidth. The benefit of the model above compared with classical seasonal adjustment processes is that the seasonal identification is not based rigidly on the calendar date, but is identified based on the observed behavior of the data through the parameters $c_{10}$ and $c_{11}$. The problem which arises is that even though the estimated start times and periodicity of the seasonal component are learned through model estimation, once it has been estimated, it is still rigidly applied over all time periods. This means that although the best ‘average’ seasonal fit is chosen, the adjustment mechanism is still unable to account for the natural stochastic variation in the placement and length of the seasonal component from year to year.

Accounting for seasonal patterns in syndromic surveillance data for outbreak detection (Burr et al., 2006) suggests a hierarchical seasonal adjustment process which updates the estimate for the seasonal component for the current time period based on learning from past observations. Specifically, the authors specify the model:

$$E(C_d) = b_y(d) + \frac{a_y}{\sigma_y} \phi \left( \frac{d - \bar{d}_y}{\sigma_y} \right) + \sum_{i=1}^{7} c_i I_i(d)$$

Where:

$$b_y(d) = b_{y-1} + \frac{d}{365} (b_y - b_{y-1})$$ accounts for the changing baseline by interpolating between the previous baseline, $b_y$ and the previous one, $b_{y-1}$.

$a_y$ is the scaled peak amplitude for year $y$

$\phi$ is the probability density function of a normal variable
and where $\Delta_y$ is the time period of the peak and $\sigma_y$ designates the duration of the peak in year $y$.

$$\sum_{i=1}^{7} c_i l_i(d)$$ captures the day of the week effects, although this component would probably be unnecessary for animal surveillance unless a weekly pattern is identified. This effect is necessarily included when dealing with human data because people tend to check themselves into hospitals in higher numbers on Mondays and Fridays. The model above was shown by the authors to improve seasonal adjustment accuracy compared with the non-hierarchical model shown originally.

The seasonal adjustment mechanisms outlined here are by no means an exhaustive list of the potential methods presented in the available literature, but they are meant to serve as a starting point to those interested in pursuing the topic further as most of the current models are separated by whether they are based on a hierarchical or non-hierarchical adjustment process. The important caveat to take away from these examples of seasonal adjustment processes are that modeling the seasonality of epidemiological data is an important part of creating an optimal data set on which to project an aberration detection algorithm. The analysis performed in this paper focuses mainly on the aberration detection algorithms themselves and methods for their improvement, without extensive analysis into the seasonal component. For additional reading regarding the seasonal adjustment process for epidemiological data, see *Automated time series forecasting for biosurveillance* (Burkom and Murphy, 2007).
CHAPTER 4

DATA ANALYSIS (DATA SET #1)

4.1 Making the case for simulated data

The data used to conduct the analytical portion of this thesis were obtained from publicly available CDC Monte Carlo simulation data which were used in the article Comparing Aberration Detection Methods with Simulated Data (Hutwagner et al., 2005) mentioned earlier. The authors’ argument for the practical relevance of detection algorithms applied to simulated data, which is repeated here, is that when working with real world data, it is often exorbitantly expensive if not impossible to investigate the exact starting and ending dates of an outbreak. This makes it infeasible to measure such things as sensitivity and mean time to detection, which are important and useful metrics when comparing aberration detection model performance. Therefore, artificial data provide a cost-effective means by which to compare different detection methods, under the presumption that models with generally better performance on simulated data will also be superior on real data, even if key performance metrics necessary to confirm this are no longer measurable.

4.11 Description of simulated data sets

The full set of simulated data contain 50 scenarios of possible baseline data behavior, each with differing parameter vectors used to simulate seasonality, trend, mean value, and variance, and whose probability density functions are distributed as negative binomial distributions. Each iteration of a baseline series, given an underlying set of generating parameters, spanned a six year period from Jan 1st, 1994 to Dec 31st, 1999. Without changing the underlying generating parameters, this six-year period was reconstructed for 10,000 total iterations. In this analysis, model comparisons are made using two different baselines, which are
identical in the mean and variance of their generating parameters, but different in their levels of seasonality. This was done specifically so that the performance of the pooling methodology could be compared to cases where seasonality is present and where it is not.

4.12 Superimposition of outbreaks

The superimposition of the outbreaks in the baseline data were conducted in the following manner (excerpt taken from Hutwagner et al., 2005):

“Ten types of outbreaks were randomly placed throughout the data streams. Days for the start of outbreaks were randomly selected using a binomial distribution; one (1) indicated the start of an outbreak and zero (0) indicated no start of an outbreak. The type of outbreak that started on a given day was selected using the remainder obtained by dividing the Julian date (1 - 365) by 10. Outbreaks did not overlap; the smallest allowed time interval from the end of one outbreak to the start of the next was 5 days.

Outbreak types one and two were 1-day spikes of magnitude equal to two and three times the standard deviations of the de-trended data, respectively. Outbreak types seven through nine and zero were based on a log normal distribution using two means, each at two and three times the de-trended standard deviations. These two means were chosen to represent explosive and more gradual outbreaks in which 95% of the cases appear from 1-4 days and from 7-13 days after exposure, respectively. In outbreak types three through six, the lognormal distribution of attributable symptomatic cases was reversed.

The lognormal distribution was selected to model the majority of the outbreaks based on work by Sartwell in 1949 and widely used since. The scenario underlying the use of this distribution is a single-source, common-vehicle outbreak, presumably resulting from a
bioterrorist attack. The use of the single-day spike signals and the reversed lognormal allows for alternative outbreak scenarios.”

4.2 Corrective data transformations

The first data set considered for analysis contained a mild trend component and no seasonality. The data were first transformed by taking the square root of the dependent variable, totalcount, which represents the total number of cases reported for a given day. This is done to induce desirable statistical properties based on the established result that the square root of a Poisson random variable is approximately normal with a standard deviation of .5 so long as the mean is sufficiently large (Johnson, Kotz, and Kemp, 1992) The obvious benefit of this procedure is that the original series need not be normal for this transformation to induce approximate normality in the new series. The following graph shows the total count square root (tcsqrt) transformed variable:

Figure 4.1: Dependent variable square root transformation
The presence of trend is clearly visible upon inspection of figure 4.1. Performing a regression on the dependent variable against time allows for the removal of the long-term trend component. Figure 4.2 shows the series with the trend component removed.

Finally, it is known that the simulated data include day-of-the-week effects, so to control for this, a 7-day moving average is calculated to eliminate such effects from finding their way into the error component of the model. The following figure represents the final series used for analysis, absent of trend or day of the week effects. It should be noted that six observations are lost due to this second transformation procedure.

**Figure 4.2 Transformed dependent variable de-trended**

```
Figure 4.3 Transformed dependent variable de-cycled
```

```
Figure 4.3 Transformed dependent variable de-cycled
```

```
Figure 4.3 Transformed dependent variable de-cycled
```
4.21 Establishing a starting point

The first step of the data analysis was to determine the smoothing parameter which minimizes the AICC\(^5\), which takes in to account the trade-off in gains to model accuracy relative to the number of parameters used in estimation. The smoothing parameter which optimizes the AICC was found to be .005721 for this data set, which means that .57% of the data, or 12 data points in the local neighborhood around the point being estimated are used in the local regression procedure. As can be seen from figure 4.4, this results in the LOESS procedure almost completely interpolating the observed data.

Figure 4.4 AICc\(_1\) optimized smoothing parameter vs dependent variable

\[ AICc_1 = n \log(\hat{\sigma}^2) + n \left( \frac{\varphi_1}{\varphi_2 - 2} \right) \]
Along with point estimates, a one-sided 99% confidence interval was also constructed; the confidence interval is one sided because we are never concerned about aberrantly low values of the dependent variable, only high ones. The model flags an observation as abnormal high if it is larger than the upper limit of the confidence interval constructed around the point estimate for that period in time. This methodology resulted in a sensitivity of 57.6%, an average time to detection of 3.95 days and standard deviation of nearly 2 days, and a specificity of 86%. These results are indicative of the problem analysts face in reality—in order to keep the number of false-positives down, the sensitivity and ability for a model to detect outbreaks is often depressed to completely unacceptable levels, as was the case here. In this instance, 42.4% of the outbreaks which occurred went undetected, and even when an outbreak was identified, it took almost four days on average for it to be discovered, and sometimes more than a week. Up until now, when faced with this dilemma the only real option was to increase the model sensitivity by decreasing the width of the confidence intervals, and in so doing drive the false-positive rate upward precipitously. It has been suggested that the optimal solution to this quandary is to focus on minimizing the cost of investigating false alarms, which upon further consideration seems to be a dangerous path to tread—if one aggressively cuts costs in order to limit the financial waste of false alarm investigations, one must consider at what point cutting costs become synonymous with cutting corners, jeopardizing the assumption that if an outbreak is present it will always be discovered upon investigation due to the desensitization of the monitoring agents (Stoto, 2007). When dealing with confidence intervals, one must always perform the due diligence of checking the model’s residuals to look for any abnormal behavior in the error structure.
Figure 4.5 Optimized smoothing parameter model residual diagnostics

Figure 4.5 indicates that the error structure for the model residuals are approximately normally distributed with mean zero, and are homoscedastic, implying the confidence interval used for aberration detection is statistically accurate.

4.22: Model selection criteria for optimal pooling candidates

In order to begin addressing an alternative solution to this problem, a new measurement must be constructed with which to compare models. This measure is the ratio of ‘hits’, defined as period in which the aberration detection model accurately identifies an outbreak, to the overall number of false positives. This new ratio will be called the, “Model Efficiency”, because it represents the amount of correct flags the model makes relative to the number of incorrect ones, which are costly to investigate. This method differs from the sensitivity metric in the crucial sense that it takes in to account the overall number of times the model’s response was correct, not simply the binary outcome of whether an outlier caused by an outbreak was identified at least once during the period the outbreak was present. This metric is quick to calculate and provides a common measuring stick by which to compare the overall accuracy of competing models. In this case, the competing models are the fitted LOESS point estimates resulting from changing the value for the smoothing parameter over an interval. The
larger the smoothing parameter, the less the LOESS estimates will interpolate the observed data, but in the process sensitivity will suffer as the point estimates are drawn closer to the overall sample mean and the confidence intervals become narrower due to the increased number of observations used in the localized regression neighborhood. The AICc optimized smoothing parameter suggests a starting point for where the interval of values to be tested for the smoothing parameter should be placed. Table 4.6a illustrates the behavior of the “Model Efficiency” ratio as the smoothing parameter is increased from .01 to .1 by hundredth increments.

Table 4.6a Efficiency frontier data table

<table>
<thead>
<tr>
<th>Smoothing Parameter</th>
<th>Hit</th>
<th>Hit%</th>
<th>Miss</th>
<th>Miss%</th>
<th>False Positive</th>
<th>False Positive %</th>
<th>Hit/False Positive Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.005721</td>
<td>12</td>
<td>5.687204</td>
<td>199</td>
<td>94.3128</td>
<td>117</td>
<td>5.357142857</td>
<td>0.102564103</td>
</tr>
<tr>
<td>0.01</td>
<td>72</td>
<td>34.12322</td>
<td>139</td>
<td>65.87678</td>
<td>225</td>
<td>10.3021978</td>
<td>0.32</td>
</tr>
<tr>
<td>0.02</td>
<td>143</td>
<td>67.77251</td>
<td>68</td>
<td>32.22749</td>
<td>347</td>
<td>15.88827839</td>
<td>0.412103746</td>
</tr>
<tr>
<td>0.03</td>
<td>169</td>
<td>80.09479</td>
<td>42</td>
<td>19.90521</td>
<td>425</td>
<td>19.45970696</td>
<td>0.397647059</td>
</tr>
<tr>
<td>0.04</td>
<td>179</td>
<td>84.83412</td>
<td>32</td>
<td>15.16588</td>
<td>611</td>
<td>27.97619048</td>
<td>0.292962357</td>
</tr>
<tr>
<td>0.05</td>
<td>189</td>
<td>89.57346</td>
<td>22</td>
<td>10.42654</td>
<td>500</td>
<td>22.89377289</td>
<td>0.378</td>
</tr>
<tr>
<td>0.06</td>
<td>195</td>
<td>92.41706</td>
<td>16</td>
<td>7.582938</td>
<td>528</td>
<td>24.17582418</td>
<td>0.369318182</td>
</tr>
<tr>
<td>0.07</td>
<td>195</td>
<td>92.41706</td>
<td>16</td>
<td>7.582938</td>
<td>541</td>
<td>24.77106227</td>
<td>0.360443623</td>
</tr>
<tr>
<td>0.08</td>
<td>196</td>
<td>92.891</td>
<td>15</td>
<td>7.109005</td>
<td>565</td>
<td>25.86996337</td>
<td>0.346902655</td>
</tr>
<tr>
<td>0.09</td>
<td>197</td>
<td>93.36493</td>
<td>14</td>
<td>6.635071</td>
<td>593</td>
<td>27.15201465</td>
<td>0.332209106</td>
</tr>
<tr>
<td>0.1</td>
<td>197</td>
<td>93.36493</td>
<td>14</td>
<td>6.635071</td>
<td>611</td>
<td>27.97619048</td>
<td>0.322422259</td>
</tr>
</tbody>
</table>

Figure 4.6b Efficiency frontier graph
These results indicate that the efficiency ratio is bimodal, rising to a global peak of .412 at a smoothing parameter value of .02, declining, and then rising again to a local peak of .378. The choice to include .02 smoothing parameter LOESS model is straightforward—it offers the most efficient detection results. We choose the .05 smoothing parameter LOESS model as our second choice because it represents a point at which sensitivity increases while actually lowering the number of false positives estimated compared with the .04 and .06 smoothing parameter models. The logic behind the choice of which two models to pool is based on which two out of all those being considered are most efficient. Evaluating the LOESS estimation results for smoothing parameters of .02 and .05 provides the following performance metrics:

<table>
<thead>
<tr>
<th>Smoothing Parameter</th>
<th>Sensitivity</th>
<th>Mean Days Until Detection</th>
<th>Std. Deviation</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>82%</td>
<td>2.75</td>
<td>1.83</td>
<td>75.82%</td>
</tr>
<tr>
<td>0.05</td>
<td>85%</td>
<td>1.5</td>
<td>0.96</td>
<td>72%</td>
</tr>
</tbody>
</table>

The following graphs demonstrate the decreased data interpolation of the LOESS model estimates as the smoothing parameter increases:

Figure 4.7b .02 LOESS smoothing parameter vs dependent variable
It is an important first step to once again examine the model residuals to determine the integrity of the confidence intervals used to determine the alarm threshold.

### 4.23 Residual diagnostics

![Figure 4.8a .02 LOESS smoothing parameter model residual diagnostics](chart)

- **Observed Values**
- **Time (in days)**
- **Dependent Variable**
- **.05 Smoothing Estimates**
The residuals for the models employing the smoothing parameters of .02 and .05 appear to have small to moderate heteroscedasticity at the higher end of the residual quantiles where the predicted values are larger. This is not altogether surprising because the predicted values are drawn upward by the occurrence of outbreaks; the larger smoothing parameters used for these models, relative to the AICc, optimized smoothing parameter of .005721 originally shown, means that the neighborhood used to approximate points before and after an outbreak, which will use the outbreak points as part of the estimation procedure, will be prone to making predictions that are too large. Efforts made to correct for this problem by re-estimating the weights applied through the tri-cube function were unsuccessful in alleviating this problem; investigation into the heuristics of controlling for this could be a potentially fruitful area for further research. As a result of the behavior of the model residuals, the confidence intervals are likely larger than necessary. Despite this, the model pooling procedure outlined below uses only the point estimates from each model, and is not influenced by their variances. The accuracy of
the model efficiency frontier, derived from graphing the observed hit/false positive ratio for different values of the smoothing parameter, would be the main beneficiary of such efforts.

Once again, we would have ended up with two competing models in which either sensitivity or specificity must be arbitrarily discounted in order to form a decision as to which model should be selected as the aberration detection algorithm for the surveillance system in question—until now.

4.3 Dynamic pooling methodology

The idea for this model pooling methodology was the result of much reflection on how to combine model estimates in such a way as to retain their positive characteristics while discarding, or at least mitigating, their negative ones. The positive characteristics of the model with a smaller smoothing parameter is that it is less prone to false-positives because the smaller neighborhoods used in the locally weighted regression process allow for a greater degree of interpolation of the observations in the series, yet this is also a double-edged sword in that greater data interpolation leads to confidence intervals which often contain genuinely abnormal values caused by outbreaks and thus go undetected. The opposite is true of the model with a larger smoothing parameter—it does not interpolate the data enough, pulling its estimates closer to the overall sample mean and resulting in confidence intervals that often indicate outbreaks where there actually are none. This is essentially the problem shown in the application of the C1, C2, and C3 detection algorithms, but applied in the context of locally weighted regressions.

The challenge then is to find some combination of the models such that the pooled model improves at least one performance metric significantly while making other performance metrics either no worse off or only marginally so relative to the gains in another area. The first
The step in accomplishing this is to apply the following pooling formula to the model predicted values:

\[ X_p = \sum_{j=1}^{m} (X_{ij} \times (1 - \frac{|X_{ij} - Y_i|}{\sum_{j=1}^{m}|X_{ij} - Y_i|})) \]

Where \(X_{ij}\) is the \(i\)th prediction by the \(j\)th model, \(Y_i\) is the observed value of the dependant variable at period \(i\), and \(X_p\) is the pooled prediction estimate. This simple formula dynamically creates a unique weight for each of the model predictions based on which model was relatively closer to the actual observed value for any given time period. For example, if model A predicts a value of 5 and model B predicts a value of 0, and the actual value is 4, the pooling methodology would result in a pooled estimate of \(5 \times (1 - \frac{1}{5}) + 0 \times (1 - \frac{4}{5}) = 4\).

Applying this transformation results in a pooled prediction series which lies somewhere between the extremes of over-interpolation and under-interpolation found in the models individually. Figure 4.9a illustrates the behavior of the new pooled prediction series vis-à-vis the observed values of the dependent variable.

\[\text{Figure 4.9a Pooled prediction series vs dependent variable}\]
4.4 Post-pooling adjustments

The residuals for the pooled model appear to be approximately normal, with a mean of 0.0546 and a standard deviation of 0.70. The larger number of outliers in the right tail of the distribution is reflective of the nature of the data set and the superimposition of the outbreaks.

A major consideration which must be addressed is the effect on the estimate reached by the pooling algorithm when an outbreak is present. The pooling algorithm stated above will always give a proportionately higher weight to the predicted value that is closer to the observed value—this means that during the period in which a disease outbreak is present, the pooled prediction value will be drawn closer to this value, working against any process which might be developed to establish an alarm threshold. The solution to this problem is to take the average weight applied to each model over all periods in the dynamic estimation phase, and to apply those weights uniformly to the original model prediction series—this will result in an overall fit that is roughly equivalent, but in the absence of dynamic pooling, will mean that the pooled
prediction series will not be as sensitive to interpolating extreme values during periods in which an outbreak is present.

4.5 Creating an alarm threshold

To complete the process, a decision mechanism must be developed which provides a way to develop an alarm threshold. To achieve this we perform a constrained maximization of the model accuracy, which is calculated as the ratio of number of individual time periods in which the pooled series correctly identifies an outbreak to the number of times it fails to do so (hit to miss ratio), subject to a performance requirement to be specified by the practitioner. In this case, the performance criterion employed is the overall false positive rate, or the total number of false-positives divided by the sample size. The standard deviation for the pooled prediction values is calculated, and the choice parameter which then must be optimized is the number of standard deviations uniformly placed above the pooled prediction values which maximizes the hit to miss ratio subject to the specified performance requirement, in this case the false positive ratio. This choice parameter should be between 0 and 3, and be denominated by hundredths. There is no probabilistic element to this ceiling; the variances of the models estimated individually are calculable, but those properties are lost in the pooling process. Thus, the optimization process proceeds in an iterative fashion. Specifically, let:

$$S = \text{the standard used to classify uncertain observations as anomalies, in this case the number of standard deviations above a predicted value after which the presence of an observed value will result in the observation being flagged as an outbreak}$$

$$\alpha(S) = \text{The model accuracy, measured by the number of correctly classified anomalies}$$

$$c = \text{the performance requirement, specified by the desired maximum false-positive ratio}$$
For the purposes of this research effort, the maximum acceptable false-positive rate was chosen to be 5%. The false-positive rate is a logical metric to use as a constraint because the cost-per-investigation of a falsely identified outbreak can be estimated fairly simply, and given the budget allotted for detection, it would be straightforward to determine the maximum amount of false-positives that could be investigated without overstepping predetermined financial constraints. The constrained maximization problem is then formally stated as:

Maximize: $\alpha(S^*)$ s.t. $c \leq .05$

$S$

Figure 4.10 graphically represents the behavior of overall model accuracy, measured by the hit/miss ratio, and the model’s specificity, measured by proxy through the false-positive ratio.

Figure 4.10 Accuracy vs false positives graph

The value $S^*$ which solves this constrained optimization is found to be 1.4, which results in the following measurements for sensitivity, specificity, and false-positives. The results from the previously estimated LOESS models with smoothing parameters of .02 and .05, from which this pooled estimate was derived, are included for comparison.
4.6 Initial results

Table 4.11a LOESS vs pooled model performance comparison

<table>
<thead>
<tr>
<th>Smoothing Parameter</th>
<th>Sensitivity</th>
<th>Mean Days Until Detection</th>
<th>Std. Deviation</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>82%</td>
<td>2.75</td>
<td>1.83</td>
<td>75.82%</td>
</tr>
<tr>
<td>0.05</td>
<td>85%</td>
<td>1.5</td>
<td>0.96</td>
<td>72%</td>
</tr>
<tr>
<td>Pooled</td>
<td>75.80%</td>
<td>3</td>
<td>1.7</td>
<td>95.01%</td>
</tr>
</tbody>
</table>

At first glance, this result suggests that the pooled model suffers from the same quandary of preference; a choice of sacrifice between sensitivity and specificity seems inevitable. However, nearly all of the disparity between the sensitivity of the optimally pooled methodology and the LOESS detection models can be attributed to the general inability of the optimally pooled model to identify small, one day spike outbreaks. In the context of developing syndromic surveillance systems that work well in the real world, this type of outbreak is highly unlikely to occur. Out of 33 observed outbreaks over the simulated six year period, eight of these outbreaks were one day spikes. When these one day spikes are eliminated from the sensitivity and mean time to detection measurements, the picture painted by comparing the models with one another changes substantially.

4.6.1 Results excluding one day spikes

Table 4.11b LOESS vs pooled model performance comparison sans 1-day outbreaks

<table>
<thead>
<tr>
<th>Smoothing Parameter</th>
<th>Sensitivity</th>
<th>Mean Days Until Detection</th>
<th>Std. Deviation</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>96%</td>
<td>2.91</td>
<td>1.73</td>
<td>75.82%</td>
</tr>
<tr>
<td>0.05</td>
<td>100%</td>
<td>1.58</td>
<td>1.01</td>
<td>72%</td>
</tr>
<tr>
<td>Pooled</td>
<td>96%</td>
<td>3.08</td>
<td>1.7</td>
<td>95.01%</td>
</tr>
</tbody>
</table>

When excluding from consideration the observations of one day spikes, the pooled methodology succeeds at the original goal of improving the performance metrics of at least one of the models while making it no worse off in another. This is the case when comparing the pooled model performance with that of the LOESS model utilizing a .02 smoothing parameter.
The sensitivities are identical, with very close mean times to detection and standard deviations, but crucially, the pooled model represents a 13% performance improvement in specificity as measured by the number of false positives relative to all data points in the sample. In absolute terms, this means 413 less false positives are flagged over a six year period when applying the pooled methodology compared to the LOESS detection model with a .02 smoothing parameter. This substantially increased specificity is obtained while sacrificing very little if anything in the areas of detection speed or sensitivity when outbreaks that arguably implausible to occur in the field in reality, i.e., small one day spikes, are excluded.

4.7 Simulated out-of-sample performance

The next step is to ascertain comparable model performance metrics for out of sample data, where perfect information is unavailable to either the model or the practitioner forming it. To simulate this scenario, a different iteration of the six year time period from Jan 1st, 1994 to Dec 31st, 1999 is evaluated at the optimal pooling parameters identified in the previous analysis. Specifically, LOESS models with .02 and .05 smoothing parameters will be estimated, which will then be uniformly pooled by multiplying each model’s predicted value at a point in time by .569249 (for the .02 smoothing parameter model) and .430751 (for the .05 smoothing parameter model), and then finally setting the number of standard deviations above the pooled prediction value past which observations are flagged as outbreaks to 1.4. The following table shows the results of applying the optimal results from the initial model to a completely separate and independent iteration that uses the same underlying parameter vectors for generation in the fashion stated above. The following graph shows the predicted values of the LOESS .02 and .05 smoothing models, the pooled model predictions that utilize the optimal parameters from the previous iteration, and the actual observations of the dependent variable.
Table 4.13a Simulated out-of-sample performance comparison

<table>
<thead>
<tr>
<th>Smoothing Parameter</th>
<th>Sensitivity</th>
<th>Mean Days Until Detection</th>
<th>Std. Deviation</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>85%</td>
<td>1.92</td>
<td>1.18</td>
<td>90.30%</td>
</tr>
<tr>
<td>0.05</td>
<td>94%</td>
<td>1.16</td>
<td>0.58</td>
<td>82%</td>
</tr>
<tr>
<td>Pooled</td>
<td>82%</td>
<td>1.89</td>
<td>1.19</td>
<td>96.50%</td>
</tr>
</tbody>
</table>

Table 4.13a shows the model prediction metrics with the one day outbreaks included in the measurement process. Once again, if the one day outbreaks are excluded, the pooled model estimates reveals a small sacrifice in sensitivity (1 outbreak not one day in length is missed), but leads to a moderate to significant improvement in specificity. To restate this benefit in terms of total values, the pooled prediction methodology results in 316 less false-positives than the .05 LOESS model, and 136 fewer than the .02 LOESS model. Whether the slight dip in sensitivity is justified depends ultimately on whether a higher false positive rate yields more investigation flags than is financially feasible for the overseeing agency to respond to.
4.8 Concluding results for data series #1

Overall, these results indicate that the optimized model pooling values which solve one series’ problem offer similar, and in this case even better, performance on data randomly generated from the same underlying parameters. This is significant because much of econometric time series is based on the assumption that:

“...All random variables of $Y_t$, whether they represent the past, present or future values of an economic variable, are assumed to have the same mean and the same variance” (Enders, 2004)

Therefore, in order for our results to be consistent with this theory, similar performance should be observed when taking optimal values derived from the first data set and applying them to the second because they are both generated by the same mean, variance, and probability density functions. The fact that this was the case suggests optimally derived values can and should be utilized in post-sample forecasting and that overall, the model pooling process is robust so long as the generating parameters used to estimate the model do not change significantly.
The data sets used for analysis thus far have contained mild to moderate trend and day of the week cyclicality, but have so far been absent of a significant seasonal component. There are many econometric time series which display moderate to extreme seasonality, such as seasonal flu, and the robustness of this methodology would not be complete without examining its performance under such baseline conditions. The results for the data employed in the next phase of analysis was desired to be directly comparable to the results found in the first portion with respect to whether the efficacy of the pooling methodology erodes or strengthens in the presence of seasonality. Because the parameter vectors generating the simulated data are identical in their true values for the mean, variance, and trend, the only difference from an experimental perspective is the inclusion of extreme seasonality in the data generating functions. Values for any of the data series will of course will be different every time they are simulated because individual observations are created through a random number generator, but the population parameters and their variances are known to have not changed.

5.1 Corrective data transformations

The same corrective data transformations were applied to the data series containing seasonality to reduce any non-normal behavior (square root transformation), and to extract trend and day of the week effects. Figure 5. shows the series before the final correction for seasonality was carried out.

5.2 Note on seasonal adjustment

Correcting for seasonality when dealing with observations of disease or sickness, as mentioned previously, is difficult because the beginning of the outbreak season, the length of time it lasts,
and pattern with which it rises, crests, and declines is different almost every year. Adjusting for seasonality is a simple process when dealing with series that have relatively predictable seasonal patterns, such as retail sales, gasoline consumption, etc. The method by which those types of series are pooled, where observations in January 2010 are compared with observations from January 2009, 2008, etc., doesn’t translate well to modeling disease outbreaks.

Figure 5.1 Square root of transformed, decycled dependent variable with seasonality

In fact, employing that kind of methodology will often end up distorting the seasonal pattern, for instance making a bi-modal flu season appear uni-modal, where the crest of the uni-modal model is in the trough of the true bi-modal one (Hafen et al., 2009). One way to correct for seasonality then is to simply create a large moving average, of 90 or 100 days, and to subtract this average value from the 91^{st} or 101^{st} observations, and so on and so forth. This is an example of an extremely simple kernel smoother; performing a robust investigation in to the properties and effects of seasonal adjustment for biosurveillance data is another thesis topic in and of itself. Because this thesis is primarily concerned with combining detection algorithms in order to improve aberration detection capabilities, the seasonal adjustment process employed is extremely simple. Further research into combining the pooling methodology outlined here
with different forms of more advanced seasonal adjustment methods may be an elucidating endeavor.

We chose a 100 day moving average because the seasonal component appears to have a long cycle period which lasts at least this long, as evidenced by the previous figure. Furthermore, it allows us to exercise one of the luxuries of working with a simulated data set, which is that we have a wealth of observations available, and so truncating a total of 106 values (the six come from the day of the week correction which uses a 7-day moving average centered on the 4th day) doesn’t significantly impact the calculations of the sample moments, as is the case when working with relatively few degrees of freedom. The graph below illustrates the final, seasonally corrected series used for estimation in which every variable from the 101st forward has had the average of the previous 100 observations subtracted from the observed value for that period.

Figure 5.2 Seasonally adjusted data series

The seasonal adjustment appears to have not completely removed this component from the data series; moving averages of 200 days and 300 days were also considered, but visual inspection of the adjusted series did not provide significantly different results from those shown
in the figure above. Ultimately the determination to move forward with estimation using the 100-day moving average seasonal adjustment method was made based on the fact that there is more practical use for a correction procedure which requires 3 months of data rather than one which requires 10 months if the results of the 100 day moving average seasonal adjustment procedure are considered acceptable.

5.3 Selection criterion for optimal model pooling candidates

Once again, the first step in identifying the two optimal smoothing parameters for pooled estimation begins by graphing the efficiency frontier for an interval of values for the LOESS smoothing parameter. A one-sided 99% upper limit confidence interval was used once to maintain consistency with the series analyzed previously.

Table 5.3a Efficiency frontier data table

<table>
<thead>
<tr>
<th>Smoothing Parameter</th>
<th>Hit</th>
<th>Hit%</th>
<th>Miss</th>
<th>Miss%</th>
<th>False Positive</th>
<th>False Positive %</th>
<th>Hit/False Positive Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.005721</td>
<td>5</td>
<td>2.109705</td>
<td>232</td>
<td>97.8903</td>
<td>50</td>
<td>2.289377289</td>
<td>0.1</td>
</tr>
<tr>
<td>0.01</td>
<td>5</td>
<td>2.109705</td>
<td>232</td>
<td>97.8903</td>
<td>50</td>
<td>2.289377289</td>
<td>0.1</td>
</tr>
<tr>
<td>0.02</td>
<td>204</td>
<td>96.07595</td>
<td>33</td>
<td>93.92405</td>
<td>244</td>
<td>11.17216117</td>
<td>0.836065574</td>
</tr>
<tr>
<td>0.03</td>
<td>220</td>
<td>92.827</td>
<td>17</td>
<td>7.172996</td>
<td>318</td>
<td>14.56043956</td>
<td>0.691823899</td>
</tr>
<tr>
<td>0.04</td>
<td>227</td>
<td>95.78059</td>
<td>10</td>
<td>4.219409</td>
<td>355</td>
<td>16.25457875</td>
<td>0.63943662</td>
</tr>
<tr>
<td>0.05</td>
<td>230</td>
<td>97.04641</td>
<td>7</td>
<td>2.953586</td>
<td>394</td>
<td>18.04029304</td>
<td>0.583756345</td>
</tr>
<tr>
<td>0.06</td>
<td>229</td>
<td>96.62447</td>
<td>8</td>
<td>3.375527</td>
<td>432</td>
<td>19.78021978</td>
<td>0.530092593</td>
</tr>
<tr>
<td>0.07</td>
<td>229</td>
<td>96.62447</td>
<td>8</td>
<td>3.375527</td>
<td>459</td>
<td>21.01648352</td>
<td>0.498910675</td>
</tr>
<tr>
<td>0.08</td>
<td>229</td>
<td>96.62447</td>
<td>8</td>
<td>3.375527</td>
<td>475</td>
<td>21.74080825</td>
<td>0.482105263</td>
</tr>
<tr>
<td>0.09</td>
<td>229</td>
<td>96.62447</td>
<td>8</td>
<td>3.375527</td>
<td>497</td>
<td>22.75641026</td>
<td>0.460764588</td>
</tr>
<tr>
<td>0.1</td>
<td>227</td>
<td>95.78059</td>
<td>10</td>
<td>4.219409</td>
<td>497</td>
<td>22.75641026</td>
<td>0.456740443</td>
</tr>
</tbody>
</table>
Figure 5.3b Efficiency frontier graph

The efficiency frontier for the seasonally adjusted model is uni-modal, with a peak at .02. This presents something of a problem because there is no clear choice for the second smoothing parameter that should be estimated to begin the model integration process. The following figure represents the actual table of values used to create the efficiency frontier for this data set, which provides some greater detail about the performance of the different smoothing parameters.

Although the hit-to-false positive ratio peaks and then declines, the number of hits produced overall by the models follows a different pattern—the model with the highest efficiency rating is not the model with the most overall hits. In the previous data set, the number of hits was a fairly linear function of the smoothing parameter—the larger the smoothing parameter, the more total correct identifications (and false-positives) were made by the model. That pattern differs here, most likely due to the seasonality still present in the data. As the smoothing parameter approaches .1 the number of hits actually decreases, reaching a peak of 230 when the smoothing parameter is .05. This is something of a sticky situation which
requires the use of logic and intuition on the part of the practitioner. In this instance, the second smoothing parameter to be selected was .05 because it maximizes the overall number of hits provided by any of the smoothing parameters considered, and thus provides the maximum sensitivity achievable compared to the alternative smoothing parameters, which have higher false positive rates and fewer total correctly identified outbreaks. The following graph shows the LOESS .02 and .05 smoothing parameter prediction results against the actual observations for the dependent variable, followed by the residuals diagnostics for each of the smoothing parameters.

Figure 5.4 .02 and .05 LOESS model predictions vs dependent variable
5.4 Residual diagnostics

Figure 5.5a .02 LOESS model residual diagnostics

Figure 5.5b .05 LOESS model residual diagnostics
These graphs show significant heteroscedasticity for the residuals throughout the entirety of the data set for both smoothing parameters, with the .05 smoothing parameter exhibiting slightly greater non-normality. This is most likely caused by the shape of the data series itself—ultimately it does not impact the ability of the pooling methodology to provide considerable performance gains, but it is definitely something which would benefit from further research. The confidence intervals used when determining which two models to pool are suspect, and different optimal model selections would likely result if residuals could be achieved that exhibit approximate normality. As done with the previous data set, the models are dynamically pooled according to the equation

\[ X_p = \sum_{j=1}^{m} (X_{ij} * (1 - \frac{|X_{ij} - Y_i|}{\sum_{j=1}^{m} |X_{ij} - Y_j|})). \]

The average weight given to each model over the entirety of the data set is then applied uniformly to the original predicted values to create the pooled model predictions. A graph of the pooled predictions against the observed dependant variable and the resulting residuals are shown below.

Figure 5.6 Pooled prediction series vs dependent variable
The pooled model residuals are approximately normal, with a mean of .23, a mode of 0, and a standard deviation of 1.3. The average weight given to the LOESS .02 smoothing parameter model was found to be .587092, and the average weight to the LOESS .05 smoothing parameter model was .412908. The residuals still are not completely clean; this is because the seasonal process is simulated and predictable—if seasonality were corrected for in the classical way, by using disjointed monthly averages and correcting for a given month by subtracting the average seasonal influence from each observation, most of the issues with non-normality within the residuals would be alleviated. The 100 day moving average method was retained on the basis that this method, although crude, is a better approximation of what a seasonal correction mechanism would look like in reality where the behavior of the seasonal component is highly variable and unpredictable. Ultimately this choice lends itself to showcasing the robustness of
local regressions and to the variety of data to which this general methodology can be successfully applied.

Figure 5.8 Model accuracy and false-positive rates

The constrained maximization presented earlier is then recalculated using the observations from the seasonal data series:

\[
\text{Maximize: } \alpha(S^*) \text{ s.t. } c \leq 0.05
\]

The performance results for the pooled model and the individually estimated LOESS models are presented below.

### 5.5 Initial results for data series #2

<table>
<thead>
<tr>
<th>Smoothing Parameter</th>
<th>Sensitivity</th>
<th>Mean Days Until Detection</th>
<th>Std. Deviation</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>89%</td>
<td>1.22</td>
<td>0.6</td>
<td>84.00%</td>
</tr>
<tr>
<td>0.05</td>
<td>92%</td>
<td>1</td>
<td>0</td>
<td>81%</td>
</tr>
<tr>
<td>Pooled</td>
<td>83%</td>
<td>1.17</td>
<td>0.46</td>
<td>95.00%</td>
</tr>
</tbody>
</table>

Out of 36 total outbreaks present in the data set, 6 are one day outbreaks. The optimally pooled model continues to have difficulty identifying one day spikes compared to the LOESS
models, however when one day outbreaks are again excluded from consideration, the optimally pooled model offers strictly better performance across all metrics save for the mean days to detection measurement, in which it is only slightly behind the .05 smoothing parameter LOESS model.

5.51 Results excluding one day outbreaks

Table 5.10 Pooled model vs LOESS performance comparison sans 1-day outbreaks

<table>
<thead>
<tr>
<th>Smoothing Parameter</th>
<th>Sensitivity</th>
<th>Mean Days Until Detection</th>
<th>Std. Deviation</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>100%</td>
<td>1.23</td>
<td>0.63</td>
<td>84.00%</td>
</tr>
<tr>
<td>0.05</td>
<td>100%</td>
<td>1</td>
<td>0</td>
<td>81%</td>
</tr>
<tr>
<td>Pooled</td>
<td>100%</td>
<td>1.17</td>
<td>0.46</td>
<td>95.00%</td>
</tr>
</tbody>
</table>

The sensitivity of the pooled and LOESS models are identical, and although having a slightly larger mean time to detection, the pooled model exhibits a sizable performance improvement with regard to specificity; an improvement of 244 fewer false-positives than the .05 LOESS model, and 141 fewer than the .02 LOESS model. If given a choice between the models above, based on the performance seen here, the dominant model is clearly the optimally pooled one. As with the originally analyzed data set, it is important to establish whether or not the pooling methodology remains robust when the optimal values derived from one data set are applied to another simulated from identical parameter vectors. This proved to be the case when seasonality was not present; we must see investigate if this still holds. If not, then there is not much use for a modeling methodology which can only identify past outbreaks well, but not future ones.
5.6 Simulated out-of-sample model performance

Figure 5.11 Simulated out-of-sample series predictions

<table>
<thead>
<tr>
<th>Smoothing Parameter</th>
<th>Sensitivity</th>
<th>Mean Days Until Detection</th>
<th>Std. Deviation</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>97%</td>
<td>1.48</td>
<td>0.94</td>
<td>88.00%</td>
</tr>
<tr>
<td>0.05</td>
<td>97%</td>
<td>1.12</td>
<td>0.55</td>
<td>81%</td>
</tr>
<tr>
<td>Pooled</td>
<td>91%</td>
<td>1.48</td>
<td>1</td>
<td>94.70%</td>
</tr>
</tbody>
</table>

LOESS models with smoothing parameters of .02 and .05 were once again estimated, and pooled predictions were formed by combining the model estimates for each period uniformly by the average weights assigned in the previous data set; in this case, .587 for the .02 smoothing parameter LOESS model and the .413 for the .05 smoothing parameter LOESS model. Finally, the upper limit threshold was set at .31 standard deviations of the pooled model predicted values above a given observation, past which point an outbreak was flagged. These results are similar to those seen previously; the pooled model has maintains the improvements seen in the in-sample results with respect to specificity while sacrificing little or no performance in the sensitivity and mean time to detection metrics, especially compared with the .02 smoothing parameter LOESS model. The data set analyzed here contained a total of 34 outbreaks, 4 of which were one-day spikes. If one day spikes are once again not considered, all
of the models achieve a 100% sensitivity rating, with average pooled model detection time being identical to that of the .02 smoothing parameter LOESS model. Once again, the optimal model pooling procedure has yielded a third choice which shows significantly better performance in one category vis-à-vis one of the two original models while not sacrificing performance in other areas. This translates to the pooled prediction series achieving comparable performance while identifying 140 fewer false-positives than the LOESS .02 smoothing parameter model, and 296 fewer false positives than the LOESS .05 smoothing parameter model.
CHAPTER 6

CONCLUSION

This research effort began with the goal of finding a simple, practical way in which practitioners looking to deploy syndromic surveillance systems could find an alternative to choosing between improved sensitivity at the cost of specificity, or vice versa. The standard by which this goal would be achieved was determined to be the development of a new model which displayed significantly improved performance in at least one of the three key model metrics (sensitivity, specificity, or mean time to detection), while becoming no worse off in the others.

To do this, we first started by comparing the model performance of various Locally Weighted Scatterplot Smoother (LOESS) models utilizing different smoothing parameters, beginning with the smoothing parameter which minimized the modified Akaike Information Criterion as a starting point. We then created an efficiency frontier which visually displayed the ratio of correctly identified outbreak periods to the number of false-positives, or periods incorrectly flagged as outbreaks when none were present. This frontier was used to determine which two models would be ideal candidates for the pooling procedure; this determination was made on the basis of which two models proved to be either local or global performance peaks for accuracy (hit to miss ratio) or efficiency (hit to false-positive ratio).

Once the two optimal candidate models were identified, the pooling procedure was conducted by applying the formula $X_p = \sum_{j=1}^{m} (X_{ij} \ast (1 - \frac{|X_{ij}-Y_{ij}|}{\sum_{j=1}^{m}(|X_{ij}-Y_{ij}|)})$). Recognizing that this formula would lead to pooled estimates that would be drawn especially close to outliers during outbreak periods, the weight given to each model’s predictions over all time periods was
averaged, and these average weights were then uniformly applied to the originally estimated LOESS model predictions.

Finally, an alarm threshold was constructed by determining the number of standard deviations, calculated from the pooled model predicted values, above each pooled model point estimate that maximized model accuracy (hit-to-miss ratio) subject to a performance requirement. The performance requirement was chosen to be a false-positive rate of no greater than 5%, although this number could be altered depending on the needs of the practitioner.

In all instances, the pooled prediction detection algorithm showed greatly improved sensitivity while maintaining no worse, or slightly better, mean time to detection than the .02 LOESS smoothing model. Sensitivity for the pooled model was at first glance consistently lower than LOESS with .02 or .05 smoothing parameters—this was shown to be almost entirely caused by the inability of the pooled model to detect one day outbreaks. When one day outbreaks were excluded from the performance calculations, in all cases, the pooled model detection methodology proved to be superior to both LOESS models in terms of specificity and on par with at least one of them in both sensitivity and average time until detection.

Recognizing that these results were tailored to retroactively optimize the models’ ability to detect past outbreaks, the optimal values for the average smoothing weights and the number of standard deviations above point estimates to set the alarm threshold were applied to completely independent six year iterations that utilized the same underlying generating parameters. This was done to approximate how the pooled models could be expected to perform when making future, out of sample predictions. These results were very encouraging in that applying previously derived optimal solutions to new data sets yielded nearly exactly the same results. This suggests that a model which optimizes the ability to detect past outbreaks.
also maximizes its chance to detect future outbreaks so long as the population parameters underlying the data generation process do not change substantially.

One area for improvement which was not addressed in this thesis is the behavior of the LOESS model residuals, which in both cases were not normal, especially so for the series containing seasonality. This heteroscedasticity more than likely resulted in inaccurate confidence intervals that were inflated, and as model performance utilizing a 99% one-sided upper limit confidence interval was the method by which the two models were selected for pooling based on their outbreak detection performance and false-positives, correcting for heteroscedasticity in the errors could very well lead to different choices of which models and their associated smoothing parameters are ideally suited to be pooled.

If the presence of one day spikes is excluded from consideration, the goal of the research effort was achieved. A third candidate model, created through a dynamic pooling procedure, was presented that provided greatly improved specificity over either of the individual models that it was estimated from, and showed little or no performance loss when compared with at least one of those models, in this case the LOESS .02 smoothing parameter model. This result, combined with the fact that the performance gains were maintained when values optimally derived from one data set were applied to another, suggest that a feasible alternative has been provided to practitioners wishing to employ syndromic surveillance systems who are not content with the dichotomy of being forced to choose strictly between sensitivity or specificity.


