Assessing North American Influenza Dynamics with Hierarchical Spatio-Temporal Models

Jessica Anderson
Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/gradreports
Part of the Statistics and Probability Commons

Recommended Citation
https://digitalcommons.usu.edu/gradreports/1214

This Report is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Plan B and other Reports by an authorized administrator of DigitalCommons@USU. For more information, please contact digitalcommons@usu.edu.
ASSESSING NORTH AMERICAN INFLUENZA DYNAMICS
WITH HIERARCHICAL SPATIO-TEMPORAL MODELS

by

Jessica Anderson

A report submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE
in Statistics

Approved:

Dr. Mevin B. Hooten
Major Professor

Dr. John R. Stevens
Committee Member

Dr. Jürgen Symanzik
Committee Member

UTAH STATE UNIVERSITY
Logan, Utah
2010
We present a general statistical modeling framework to characterize continental-level influenza dynamics in the United States for the purposes of examining state-level epidemiological sources and sinks. The methods we describe depend directly on state-level influenza data that are updated weekly and available on the internet. Advances in search engine query analysis have provided powerful new tools for collecting epidemiological data and, when used in conjunction with sophisticated statistical models, allow for the identification and quantification of the flow of influenza across the continental United States. Our proposed methods, when conditioned on this comprehensive search query product, can provide unprecedented scientific learning about large-scale pathways and barriers to disease transmission which can ultimately be helpful for policy, remediation, and response efforts.
ACKNOWLEDGMENTS

I would like to acknowledge my advisor Mevin Hooten. He has been a great support and an amazing resource during my time as his student. Also, I thank Lance Waller for his insight and expertise in the epidemiology aspect of this project. I would also like to acknowledge Rajan Patel and Google for their assistance and funding for this project.

I would like to acknowledge the support of my husband, Mark Anderson, and my family. Thank you for your help and support in my life. Finally, I would like to thank my God for giving me knowledge and understanding.

Jessica Anderson
# CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vi</td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Influenza Background</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Modeling Spread of Disease</td>
<td>2</td>
</tr>
<tr>
<td>1.2.1 Epidemiological Models</td>
<td>2</td>
</tr>
<tr>
<td>1.2.2 Hierarchical Models</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Influenza-Like Illnesses</td>
<td>5</td>
</tr>
<tr>
<td>2 EXPLORATORY ANALYSIS</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Analysis of Spatial Structure</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Empirical Orthogonal Functions</td>
<td>14</td>
</tr>
<tr>
<td>3 MATERIALS AND METHODS</td>
<td>16</td>
</tr>
<tr>
<td>3.1 Model Construction</td>
<td>16</td>
</tr>
<tr>
<td>3.2 Model Fitting</td>
<td>18</td>
</tr>
<tr>
<td>3.3 Model Implementation</td>
<td>18</td>
</tr>
<tr>
<td>3.4 Assessing Model Fit</td>
<td>22</td>
</tr>
<tr>
<td>3.5 Covariate Selection</td>
<td>23</td>
</tr>
<tr>
<td>4 RESULTS</td>
<td>26</td>
</tr>
<tr>
<td>4.1 Model Selection</td>
<td>26</td>
</tr>
<tr>
<td>4.2 Model Results</td>
<td>26</td>
</tr>
<tr>
<td>5 DISCUSSION</td>
<td>31</td>
</tr>
<tr>
<td>6 CONCLUSION</td>
<td>33</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>35</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Using deviance for model selection</td>
<td>26</td>
</tr>
<tr>
<td>4.2 Posterior means (and 95% credible intervals) for covariate coefficient marginal posterior distributions</td>
<td>27</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure | Page
---|---
2.1 (a) GFT time-series data showing total infections per 100,000, (b) Moran’s I statistic over the study period, and (c) p-values corresponding to the Moran’s I statistics. (Note that the original GFT data occur at weekly intervals, but we are indexing the x-axis by month for clarity.) | 10
2.2 Moran’s I test for spatial dependence on the residuals of an AR[1] model over the study period (note that the original GFT data occur at weekly intervals, but we are indexing the x-axis by month for clarity): (a) Moran’s I statistics, (b) corresponding p-values, and (c) coefficient estimates. | 11
2.3 Moran’s I test for spatial dependence on the residuals of a model regressed on state centroid locations over the study period (note that the original GFT data occur at weekly intervals, but we are indexing the x-axis by month for clarity): (a) Moran’s I statistic, (b) corresponding p-values, (c) coefficient estimates for latitude, and (d) coefficient estimates for longitude. | 12
2.4 Moran’s I test for spatial dependence on the residuals of a model regressed on a measure of population density (state population/state area) over the study period (note that the original GFT data occur at weekly intervals, but we are indexing the x-axis by month for clarity): (a) Moran’s I statistic, (b) corresponding p-values, (c) coefficient estimates. | 13
2.5 First and second empirical orthogonal functions for the GFT data. (Note that the original GFT data occur at weekly intervals, but we are indexing the x-axis for the scores by month for clarity.) | 15
3.1 Covariate maps: (a) population density (note that darker shades indicate higher population density), (b) state area (note that darker shades indicate larger area), (c) minimum January temperature (note that darker shades indicate warmer January temperatures), (d) minimum July temperature (note that darker shades indicate warmer July temperatures), and (e) age sensitivity (note that darker shades indicate increased age sensitivity).  

4.1 MCMC samples from the distributions of (a) the intrastate covariate coefficients (β_W), (b) the interstate covariate coefficients (β_B), (c) the air covariate coefficient (β_A), and (d) σ^2. For (a-b), black represents the coefficient for area, red represents the coefficient for PPSM, green represents the coefficient for minimum January temperature, and blue represents the coefficient for minimum July temperature. In (d) the square root of the samples is displayed for simplicity.  

4.2 Transmission surfaces: (a) flow surface of interstate transmission, where ILI transmits more easily between lighter states, and (b) intrastate transmission, where ILI increases more easily in darker states.  

4.3 Map of susceptibility of state i to ILI from state j, where darker states are more susceptible to obtaining ILI from surrounding states.  

4.4 The Utah example of susceptible, infected, and recovered curves over time, where N represents the total population of Utah estimated using the 2000 census. (Note that the original GFT data occur at weekly intervals, but we are indexing the x-axis by month for clarity.)
1.1 Influenza Background

Influenza is a prominent topic for most humans, as it affects many lives on a regular basis. Every winter a different strain of influenza, a mutation distinct from previous years of influenza, sweeps through populations and infects people, causing symptoms including, but not limited to, headache, fever, chills, aches, and nausea, according to the Centers for Disease Control and Prevention (CDC) Web site (Centers for Disease Control and Prevention, 2010a). When a person who is not immune comes in contact with a strain of influenza, the person may become infected. A person does not show symptoms during the incubation period (i.e., the first one to three days of being infected) and is not infectious until they begin showing symptoms (Yang et al., 2009).

Little is known about the transmission of influenza, though it is believed to happen in multiple ways (Earn et al., 2002). Droplet transmission occurs by person to person contact (e.g., an infected person coughs or sneezes on a non-infected person, who then becomes infected) (Bridges et al., 2003; Earn et al., 2002). Droplet transmission cannot alone account for the occasionally eruptive spread of influenza, which suggests some form of airborne transmission (Earn et al., 2002). Therefore, to examine the spread of influenza, it is reasonable to characterize human movement across space and throughout time, allowing for some uncertainty due to transmission type.

The spatial transmission patterns that may exist can be important in delaying the spread of influenza during a pandemic (Colizza et al., 2007; Cooper et al., 2006). The World Health Organization (WHO) Web site defines a pandemic as a “worldwide epidemic of a disease”, where an epidemic “occurs when there are more cases of that disease than normal” (World Health Organization, 2010). The WHO Web site also states that “current epidemiological models project that [an influenza] pandemic could result in two to 7.4 million deaths globally”. Influenza pandemics have occurred throughout history. In 1918, the largest recorded pandemic, commonly referred to as the Spanish flu, spread throughout
the world killing millions of people (Ghendon, 1994). There were several other pandemics during the twentieth century, though none as fierce as the Spanish flu (Ghendon, 1994). Influenza pandemics have occurred sporadically, with no recognizable pattern for how often they will occur. Most recently, the swine flu of 2009 unexpectedly moved swiftly throughout the world, killing an estimated 12,470 (Centers for Disease Control and Prevention, 2010c) and causing panic (Yang et al., 2009). Delaying a pandemic would allow more time for health organizations to prepare vaccinations for the pandemic which may appear in their region in the future (Cooper et al., 2006).

Since influenza is passed from person to person, one important factor that likely affects the way it is transmitted is population density. People living in cities (or states) with higher population densities may have a higher chance of infecting other people, because they will come in contact with more people in general (Altizer et al., 2006). It is also suggested that age may affect transmission of influenza, as elderly people and children are more likely to become infected than healthy middle-aged people (Bridges et al., 2003; Earn et al., 2002; Yang et al., 2009). Therefore, a state with a large population of children and elderly people should have a higher rate of infection. The seasonality of influenza hints that climate and weather conditions may largely affect transmission of this virus (Altizer et al., 2006; Bridges et al., 2003), though this could be indirectly related. Finally, due to modern forms of travel, it is reasonable to allow for influenza to be transmitted by passengers on aircraft (Brownstein et al., 2006; Grais et al., 2004).

1.2 Modeling Spread of Disease

1.2.1 Epidemiological Models

A traditional deterministic approach to characterizing infectious disease is with the epidemiological Susceptible, Infected, Recovered (SIR) model (Earn et al., 2002; Ma and Xia, 2009). Letting $N$ represent the total number of subjects, we divide the population into these three components such that $N = S + I + R$. This general compartment model mimics diseases like influenza where a person is susceptible ($S$) to a disease, becomes infected ($I$),
with the disease, then recovers (R) from it. In most cases, the actual counts of each of these compartments are not observed, and statistical inference is used in determining the number of people in each group (Khan et al., 2009). There are many variations to this model, including Susceptible, Infected, Susceptible (SIS), Susceptible, Infected, Recovered, Susceptible (SIRS), Susceptible, Exposed, Infected (SEI), Susceptible, Exposed, Infected, Susceptible (SEIS), and Susceptible, Exposed, Infected, Recovered (SEIR) (Ma and Xia, 2009). These compartment models have been used in explaining the rise and fall of different diseases including plague, cholera, human immunodeficiency virus (HIV), severe acute respiratory syndrome (SARS), influenza, and many others (Ma and Xia, 2009).

The SIR model was developed by Kermack and McKendrick (1927). The basic model is written as a coupled set of differential equations:

\[
\begin{align*}
\frac{dS}{dt} &= -aSI + cR, \quad (1.1) \\
\frac{dI}{dt} &= aSI - bI, \quad (1.2) \\
\frac{dR}{dt} &= bI - cR, \quad (1.3)
\end{align*}
\]

where \( S \) is the number of susceptible subjects at time \( t \), \( I \) is the number of infected subjects at time \( t \), \( R \) is the number of recovered subjects at time \( t \), \( a \) is the proportion of subjects who become infected, \( b \) is the proportion of subjects who are no longer infected and have a temporary immunity, and \( c \) is the proportion of subjects who lose immunity and again become susceptible. These models have the assumption that the disease being modeled is measured instantaneously (Ma and Xia, 2009). This is not usually the case. Also, standard implementation of these models do not take into account population or disease dynamics such as age, gender, or weather conditions that may affect the transmission of the disease. Finally, if using only the SIR models, the disease cannot be latent (i.e., the period of disease must be the same length as the period of infectiousness). An exposed (E) compartment could be added to account for latency (Ma and Xia, 2009), but this is not usually done when modeling influenza, as other factors are often much more important (Earn et al., 2002).
1.2.2 Hierarchical Models

One approach that can account for both uncertainty due to a scientific process and uncertainty due to observation error is a hierarchical (or mixed) model (e.g., Smith et al., 2002; Strefarlis and Gibson, 2002; Wikle and Hooten, 2010). In the Bayesian framework, hierarchical models can be set up using a data model which depends on an underlying process, a process model which depends on random parameters, and a parameter model where prior distributions are specified for model parameters (Berliner, 1996).

For example, suppose we have a process $\mu$ that we would like to estimate using covariates $X$ and conditioned on observed data $y = (y_1, \ldots, y_n)$. Without accounting for uncertainty due to the process, we might consider a typical regression analysis. Taking into account uncertainty due to the process, we can consider a hierarchical regression model. Assuming conditional normally distributed data $y$, the hierarchical model can be written in terms of a data model, where the $y_i$ and $\mu_i$ are independent and normal with identical variance $\sigma_y^2$ and $\sigma_{\mu}^2$, respectively,

$$y \sim \text{Normal}(\mu, \sigma_y^2 I),$$

(1.4)

a process model,

$$\mu \sim \text{Normal}(X\beta, \sigma_{\mu}^2 I),$$

(1.5)

and specified prior distributions (i.e., the parameter model),

$$\beta \sim \text{Normal}(\mu_{\beta}, \Sigma_{\beta}),$$

(1.6)

$$\sigma_y^2 \sim \text{Inv.Gamma}(r_y, q_y),$$

(1.7)

$$\sigma_{\mu}^2 \sim \text{Inv.Gamma}(r_{\mu}, q_{\mu}),$$

(1.8)

where $r_y$, $q_y$, $r_{\mu}$, and $q_{\mu}$ are strictly positive. Often, hyperpriors (i.e., fixed values) are specified for the hyperparameters (e.g., $\mu_{\beta}, \Sigma_{\beta}, r_y, q_y, r_{\mu}, q_{\mu}$) within the parameter model.

Now, we are interested in the distribution of the process and parameters given the data: $[\mu, \beta, \sigma_y^2, \sigma_{\mu}^2 | y]$ (e.g., Cressie et al., 2009). Using Bayes' Rule,

$$[\mu, \beta, \sigma_y^2, \sigma_{\mu}^2 | y] \propto [y | \mu, \sigma_y^2] [\mu | \beta, \sigma_{\mu}^2] [\beta | \sigma_y^2] [\sigma_{\mu}^2],$$

(1.9)
assuming *apriori* independence of the parameters. Thus, we are able to make inference on the coefficients, variance components, and also an underlying process of interest (e.g., Cressie et al., 2009; Hooten and Wikle, 2010).

In our case, with the focus on modeling disease, the seasonal influenza epidemic is the process of interest, and the data consist of the number of people who have influenza-like illnesses. Although we do not know the exact process of the epidemic, we are able to estimate it using the data.

1.3 Influenza-Like Illnesses

The CDC keeps track of the number of people with influenza-like illnesses (ILI) by having physicians report the number of ILI cases at their clinic on a weekly basis. ILI is defined by the CDC as “fever (temperature of 100F [37.8C] or greater) and a cough and/or a sore throat in the absence of a known cause other than influenza” (Centers for Disease Control and Prevention, 2010a). According to the CDC Web site, data collection, summary, and analysis take approximately two weeks to be processed and repeated. If a large epidemic were to occur, the public would know about it long after it was underway. Recently, a team at Google created a model that linked the number of influenza-related queries to the number of people with ILI. They tested their model against previous CDC data and concluded that they could use influenza-related queries to determine the number of people with ILI (Ginsberg et al., 2009). These Google data are also provided on a weekly basis, and are available within days of data collection (Google, 2010).
CHAPTER 2
EXPLORATORY ANALYSIS

The data used for this analysis are taken from the Google Flu Trends (GFT) project, and will be referred to hereafter as the GFT data. The number of people, per 100,000, with influenza-like illnesses are reported for each state and week on the GFT Web site (Google, 2010). Thus, the data consist of counts ranging from 0 to 100,000. The areal nature of the data may be useful for spatial analysis, as there are many approaches we can use to describe the spatial connectivity between states and its effect on influenza transmission (Bennet and Haining, 1985). Since the data are reported weekly, we could use a dynamic model that mimics an SIRS auto-regressive model which accounts for the fact that the number of people who are sick during one week is a function of the number of sick people the previous week. In what follows, we explore the GFT data to determine what type of modeling approach to take.

2.1 Analysis of Spatial Structure

The GFT data are modeled such that they have discrete spatial support, and hence it is not appropriate to analyze the data as if they were continuous in space (though the influenza epidemic process likely is). These types of data are referred to as areal or lattice data (Cressie, 1993). In the case of continuous data, the focus would be on the use of geostatistical methods, such as the variogram or covariogram, to characterize spatial autocorrelation (i.e., spatial dependence decreases as distance increases from a point of interest). The equivalent methods for describing spatial structure using areal data involve the Geary’s C and Moran’s I statistics (Schabenberger and Gotway, 2005); the exploratory analysis that follows focuses on the latter.

The Moran’s I statistic is calculated by

\[
I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (y_i - \bar{y}) (y_j - \bar{y})}{\left( \sum_{i=1}^{n} (y_i - \bar{y})^2 \right) \left( \sum_{i \neq j} w_{ij} \right)},
\]

(2.1)
where \( n \) is the number of spatial regions, \( y \) is an observation of the random variable of interest in an area denoted by \( i \) or \( j \), \( \bar{y} \) is the grand mean of all observations, and \( w_{ij} \) is the \( ij \)th element of the spatial proximity matrix, \( W \), corresponding to the connectivity of area \( i \) and area \( j \). The matrix \( W \) can be chosen in many different ways and depends largely on the situation.

A high value of \( I \) (i.e., close to positive one) suggests positive spatial autocorrelation (i.e., a spatial region having a variable of large magnitude is likely surrounded by regions having the same variable also of large magnitude), and a low value of \( I \) (i.e., close to negative one) suggests negative spatial autocorrelation (i.e., a spatial region having a variable of large magnitude is likely surrounded by regions having the same variable of small magnitude). Values of \( I \) near zero indicate no spatial autocorrelation. The expected value and variance of the Moran's I statistic are given by

\[
E(I) = -\frac{1}{n-1} \tag{2.2}
\]

and

\[
Var(I) = \frac{n^2 s_1 - n s_2 + 3 s_0^2}{(n-1)(n+1) s_0^2} - \left( \frac{1}{n-1} \right)^2, \tag{2.3}
\]

where

\[
s_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij},
\]

\[
s_1 = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left( w_{ij} + w_{ji} \right)^2,
\]

\[
s_2 = \sum_{k=1}^{n} \left( \sum_{j=1}^{n} w_{kj} + \sum_{i=1}^{n} w_{ik} \right)^2.
\]

Then the standard normal distribution can be used under the assumption of normality for testing the null hypothesis of no spatial correlation. This is done using

\[
z = \frac{I - E(I)}{\sqrt{Var(I)}}, \tag{2.4}
\]

where \( z \) is a value from the standard normal distribution under the null hypothesis. Large values of \(|z|\) indicate spatial autocorrelation.
It is often necessary to remove trends in the data before testing for spatial dependence. In these cases, the Moran’s I statistic needs modifications since the mean is heterogeneous. After removing trends in the data (e.g., by fitting a linear regression model), we can test for spatial dependence in the residuals (i.e., test the residuals of a fitted linear model for spatial dependence). Letting

\[ I_{res} = \frac{n}{\sum_{i \neq j} w_{ij}} \frac{e' We}{e'e}, \]  

(2.5)

where

\[ e = y - X\hat{\beta}, \]

and assuming normality and independent errors, \( I_{res} \) has asymptotically normal properties. The mean and variance of \( I_{res} \) are

\[ E(I_{res}) = -\frac{n}{(n - p) \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}} tr \left( (X'X)^{-1} X'WX \right), \]  

(2.6)

and

\[ Var(I_{res}) = \frac{n^2}{(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij})^2 (n - p)(n - p + 2)} \times \left[ \frac{1}{2} \sum_{i \neq j} (w_{ij} + w_{ji})^2 + 2tr(G^2) - tr(F) - \frac{2(tr(G))^2}{n - p} \right], \]  

(2.7)

where

\[ F = (X'X)^{-1} X' (W + W')^2 X, \]

\[ G = (X'X)^{-1} X' WX, \]

and \( p \) is the number of covariates plus one (Schabenberger and Gotway, 2005). Thus,

\[ z = \frac{I_{res} - E(I_{res})}{\sqrt{Var(I_{res})}} \sim N(0, 1). \]  

(2.8)

We can use these properties to test whether there is spatial correlation between neighboring areas (defined by the spatial proximity matrix \( W \)) after accounting for potential covariate effects. Large values of \(|z|\) indicate spatial correlation.
There are several ways to define the spatial proximity matrix $W$. One common way is the $k$ nearest neighbor approach. In this case, one could define

$$w_{ij} = \begin{cases} 
1, & \text{if centroid of area } j \text{ is one of } k \text{ nearest to area } i \\
0, & \text{otherwise} 
\end{cases}$$

or based on contiguity, i.e.,

$$w_{ij} = \begin{cases} 
1, & \text{if area } i \text{ and area } j \text{ have common boundary} \\
0, & \text{otherwise}. 
\end{cases}$$

A more sophisticated way of taking the second approach would be to define

$$w_{ij} = \frac{l_{ij}}{l_i}$$

where $l_{ij}$ is the length of common boundary between area $i$ and area $j$ and $l_i$ is the perimeter of area $i$. For this study, the spatial proximity matrix will be defined using the former definition of $w_{ij}$, where the centroids are the centroid locations of each state, $k = 4$, and $w_{ii} = 0$. The centroid locations are available in R in the `maps` package (R Development Core Team, 2010; Becker et al., 2010).

Figures 2.1 - 2.4 show the GFT data, Moran’s I values, associated p-values, and $\hat{\beta}$s over time for different models. Figure 2.1 (a) shows the GFT data as a set of time-series where each line represents a state, so we see the general influenza trend in time. Figure 2.1 (b) shows the Moran’s I statistic at each time point with a reference line at the expected value. Figure 2.1 (c) shows the associated p-values corresponding with Figure 2.1 (b), with a reference line at 0.05. This reference line does not necessarily show significance, as there would need to be a correction to the significance level for multiple testing. However, it is provided to generally illustrate when there might be stronger spatial structure. Figure 2.2 (a-c) displays the Moran’s I statistic and associated p-values on the residuals, and the coefficients for a first-order autoregressive linear model. We see much less spatial structure once the seasonal effect has been accounted for, but still we see potentially strong spatial structure reoccurring each year.

Figures 2.3 and 2.4 are similar to Figure 2.2, but for different models. First, the data were regressed on two location parameters: the latitude and longitude of the state centroid.
Hence, Figure 2.3 (c, d) shows coefficients for this model. Figure 2.4 is for the data regressed on a measure of population density (i.e., population/area). Again, the plots containing p-values have general reference lines at $\alpha = 0.05$ which should not be used to declare statistical significance, but rather to generally indicate strength of spatial structure.

Figure 2.1. (a) GFT time-series data showing total infections per 100,000, (b) Moran’s I statistic over the study period, and (c) p-values corresponding to the Moran’s I statistics. (Note that the original GFT data occur at weekly intervals, but we are indexing the x-axis by month for clarity.)
Figure 2.2. Moran’s I test for spatial dependence on the residuals of an AR[1] model over the study period (note that the original GFT data occur at weekly intervals, but we are indexing the x-axis by month for clarity): (a) Moran’s I statistics, (b) corresponding p-values, and (c) coefficient estimates.
Figure 2.3. Moran's I test for spatial dependence on the residuals of a model regressed on state centroid locations over the study period (note that the original GFT data occur at weekly intervals, but we are indexing the x-axis by month for clarity): (a) Moran's I statistic, (b) corresponding p-values, (c) coefficient estimates for latitude, and (d) coefficient estimates for longitude.
Figure 2.4. Moran’s I test for spatial dependence on the residuals of a model regressed on a measure of population density (state population/state area) over the study period (note that the original GFT data occur at weekly intervals, but we are indexing the x-axis by month for clarity): (a) Moran’s I statistic, (b) corresponding p-values, (c) coefficient estimates.
2.2 Empirical Orthogonal Functions

Empirical Orthogonal Functions (EOFs) are the equivalent of principal components (PCs) for spatio-temporal data. EOFs are derived in the same way as the PCs, however they account for structure both spatially and temporally (Preisendorfer, 1988). When the data can be viewed as maps of a spatial domain over time, the first EOF will be the map that explains most variation and will be accompanied by a time series indicating the time periods that the map is expressed in the data. The second EOF will be orthogonal to the first, as in principal component analysis, and will indicate the map that explains the second most variation with its accompanying time series.

Figure 2.5 shows the first two EOFs for the GFT data. Together, these two EOFs account for approximately 90% of the variation in the spatio-temporal influenza data. The first EOF shows a pattern corresponding to the seasonality of influenza. That is, influenza is more prevalent in winter months, hence the first EOF is mostly dark with large spikes in the time series during the winter months. This is to be expected, since influenza is a seasonal disease. The second EOF can be interpreted as the direction that influenza moved across the continental United States during each influenza season. For example, during the winter of 2003-2004 (the first influenza season for which we have data), the epidemic started in the west and quickly traveled to the east. The first large spike means that the EOF map was present, and the immediate spike in the opposite direction means that the exact opposite of the EOF map was present. The second influenza season (2004-2005) started in the east and spread to the west. The other seasons can be similarly interpreted.
Figure 2.5. First and second empirical orthogonal functions for the GFT data. (Note that the original GFT data occur at weekly intervals, but we are indexing the x-axis for the scores by month for clarity.)
CHAPTER 3
MATERIALS AND METHODS

3.1 Model Construction

When specifying a model for the GFT data, it is important to take into account the natural process of influenza-like illnesses (ILI). That is, a person is susceptible (S) to the disease, then infected (I) by the disease, and finally recovered (R) from the disease. After a recovery period, the person is again susceptible. This type of epidemiological model is called an SIRS model. Letting $N$ represent the total number of people, we divide the population into three components: $N = S + I + R$. Recall, the differential equations associated with this model are given in equation 1.1, and are used to model instantaneous counts for each compartment. The 2000 US Census provides a fairly recent estimate of the total population size $N$. However, we have no information for $S$, $I$, or $R$ directly. Therefore, we could use the number of people with influenza-like illnesses (ILI) from the GFT data to estimate the number of infected people. Due to the incubation period of influenza, becoming infected occurs before becoming infectious. However, this period is not often accounted for (Earn et al., 2002). So for this analysis, we transform the GFT data to represent the number of infectious people, $I$, on any given week $t$ in state $i$. A common model (Yang et al., 2009) for the natural history of influenza is that duration of infectiousness extends from 1 to 7 days with probabilities

$$ p = (1, 1, 1, 0.8, 0.6, 0.4, 0.2)' . $$

To discretize the SIRS model, we assume that ILI infection occurs uniformly throughout the week. Then, to transform the percentage of people with ILI into the number of infectious people, $I$, we have the following relationship:

$$ I_{ILI_t} = \frac{1}{N_i} \int_{t-1}^{t} \tilde{I}(\nu) \, d\nu. \quad (3.1) $$

Hence,

$$ I_t = \int_{t-1}^{t} P(\nu) \tilde{I}(\nu) \, d\nu = \frac{5}{7} N_i I_{ILI_t}. \quad (3.2) $$
Thus, the data model which is given in terms of ILI,

\[ y_{it} \sim \text{Binomial} \left( 100,000, ILI_{it} \right), \]  

(3.3)
can be reparameterized in terms of I by solving equation 3.2:

\[ y_{it} \sim \text{Binomial} \left( 100,000, I_{it} \left( \frac{7}{5N_i} \right) \right), \]  

(3.4)
where \( i = 1, \ldots, 48 \) and \( t = 1, \ldots, T \). The process model is given by:

\[ I_{it} \sim \text{Normal} \left( \mu_{it}, \sigma^2 \right), \]  

(3.5)

\[ \mu_{it} = x_i^T \beta_W I_{it-1} + \sum_{j=1}^{N_i} b_{ij} (x_j - x_i)' \beta_B I_{jt-1} + \sum_{j=1}^{n} a_{ij} \beta_A I_{jt-1}, \]  

(3.6)
where \( x_i \) and \( x_j \) are covariate values for states \( i \) and \( j \), respectively, \( b_{ij} \) is the weight of state \( j \) for state \( i \), defined by the proportion of bordering perimeter, and \( a_{ij} \) consists of the number of people flying from state \( i \) to state \( j \).

In this analysis, we are interested whether changes in influenza activity can be attributed to a state, to its neighboring states, or to air travel. The covariate coefficients, \( \beta_W \), correspond to changes in influenza activity coming from within state \( i \), the coefficients, \( \beta_B \), correspond to changes in influenza activity in state \( i \) transferring from neighboring states \( j \), and the coefficient, \( \beta_A \), corresponds to changes in influenza activity in state \( i \) transferring from state \( j \) via air travel. Thus, the parameter model is given by:

\[ \beta = (\beta_W, \beta_B, \beta_A)' \sim \text{Normal} \left( 0, \sigma^2 \beta I \right), \]  

\[ \sigma^2 \sim \text{Inv.Gamma} (r, q). \]

The posterior distribution is given by:

\[ \left[ \beta, \sigma^2, \{I_{it}\} \mid \{y_{it}\} \right] \propto \prod_{i=1}^{n} \prod_{t=1}^{T} \left[ y_{it} | I_{it} \right] \prod_{t=2}^{T} I_{t} | I_{t-1}, \beta, \sigma^2 \left[ \beta \right] \left[ \sigma^2 \right] \]  

(3.7)

Using \( I \) at each time point as the number of infectious people each week, we can also obtain the number of recovered (R) and susceptible (S) people:

\[ R_{i,t} = R_{i,t-1} + I_{i,t} \left( \frac{2}{5} \right) - cR_{i,t-1}, \]  

(3.8)
where $c$ represents the proportion of people who lose immunity and again become susceptible, and

$$S_{i,t} = N_t - I_{i,t} - R_{i,t}. \quad (3.9)$$

### 3.2 Model Fitting

The posterior distribution in equation 3.7 cannot be found analytically. However, we can use a Markov chain Monte Carlo (MCMC) algorithm to obtain random samples from the distribution of interest in order to make inference. In this analysis, the distributions of each parameter conditioned on all other parameters (i.e., the full-conditional distributions) can be found analytically. Thus, the Gibbs sampler algorithm will be used to obtain samples from the full-conditionals. Together, these distributions allow us to approximate the posterior distribution, and thus we can make the desired statistical inference on the underlying process and parameters.

To obtain samples from these distributions, we will set initial values for each of the parameters. Then, for every iteration in the MCMC algorithm, sample from each full-conditional distribution, using the sample for that parameter when sampling from the next distribution. Convergence of the MCMC algorithm occurs after a burn-in period where the parameters are searching for the stationary distribution. Thus, the first burn-in samples obtained using MCMC should be discarded before using the remainder of the samples to compute posterior quantities.

### 3.3 Model Implementation

To implement the Gibbs sampler MCMC algorithm, we need to find the full-conditional distributions. The full-conditional for $\beta$ is given by

$$[\beta | \cdot] = [\beta | \{y_{it}\}, \{I_t\}, \sigma^2]$$

$$\propto \prod_{i=2}^T [1_{I_{it}} | 1_{I_{t-1}}, \beta, \sigma^2] [\beta].$$
In this case, it helps to rewrite the model as

\[
I_t = Q_{t-1} \beta + \epsilon_t,
\]

\[
\epsilon_t \sim \text{Normal} \left( 0, \sigma^2 I \right),
\]

\[
Q = Q(I_{t-1}, X),
\]

\[
t = 2, \ldots, T,
\]

where \( Q \) is a function of the number of infectious people at the previous time and the covariates. When written this way, we have a product of three Gaussian distributions for the full-conditional distribution of \( \beta \), which will result in a Gaussian full-conditional distribution.

\[
|\beta| \propto \prod_{i=2}^{T} \left[ I_t | I_{t-1}, \beta, \sigma^2 \right] [\beta]
\]

\[
\propto \prod_{i=2}^{T} \exp \left\{ -\frac{1}{2} \left( I_t - Q_{t-1} \beta \right)^{\prime} (\sigma^2 I)^{-1} \left( I_t - Q_{t-1} \beta \right) \right\} \exp \left\{ -\frac{1}{2} \beta^{\prime} \left( \sigma^2 I \right)^{-1} \beta \right\}
\]

\[
\propto \exp \left\{ -\frac{1}{2} \left( \sum_{t=2}^{T} \left( I_t - Q_{t-1} \beta \right)^{\prime} (\sigma^2 I)^{-1} \left( I_t - Q_{t-1} \beta \right) + \beta^{\prime} \left( \sigma^2 I \right)^{-1} \beta \right) \right\}
\]

\[
\propto \exp \left\{ -\frac{1}{2} \left( -2 \sum_{t=2}^{T} \left( I_t^\prime (\sigma^2 I)^{-1} Q_{t-1} \right) \beta 
+ \beta^{\prime} \left( \sum_{t=2}^{T} Q_{t-1}^{\prime} (\sigma^2 I)^{-1} Q_{t-1} + (\sigma^2 I)^{-1} \right) \beta \right) \right\}
\]

Thus,

\[
|\beta| \sim \text{Normal} \left( A^{-1} b, A^{-1} \right)
\]

where,

\[
A = \sum_{t=2}^{T} Q_{t-1}^{\prime} (\sigma^2 I)^{-1} Q_{t-1} + (\sigma^2 I)^{-1},
\]

\[
b = \sum_{t=2}^{T} \left( I_t^\prime (\sigma^2 I)^{-1} Q_{t-1} \right).
\]
The full-conditional for $\sigma^2$, using the same reparameterized model as in the full-conditional for $\beta$, can be written as:

$$\begin{align*}
\left[\sigma^2 \mid \cdot \right] &\propto \prod_{i=2}^{T} \left[ I_t \mid I_{t-1}, \beta, \sigma^2 \right] \left[ \sigma^2 \right] \\
&\times \prod_{i=2}^{T} \left( 2\pi \right)^{-\frac{3}{2}} \left| \sigma^2 I \right|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \left( I_t - Q_{t-1} \beta \right)' \left( \sigma^2 I \right)^{-1} \left( I_t - Q_{t-1} \beta \right) \right\} \\
&\times \frac{1}{r \Gamma(q)} \sigma^{2-(q+1)} \exp \left\{ -\frac{1}{r \sigma^2} \right\} \\
&\propto \sigma^{2-\left( \frac{n(T-1)}{2} + q + 1 \right)} \exp \left\{ -\frac{1}{2 \sigma^2} \left( I_t - Q_{t-1} \beta \right)' \left( I_t - Q_{t-1} \beta \right) - \frac{1}{r} \right\} \\
&\propto \sigma^{2-\left( \frac{n(T-1)}{2} + q + 1 \right)} \exp \left\{ -\frac{1}{\sigma^2} \left( \left( \frac{1}{2} I_t - Q_{t-1} \beta \right)' \left( I_t - Q_{t-1} \beta \right) - \frac{1}{r} \right) \right\} \\
\sigma^2 \mid \cdot &\sim \operatorname{Inv.Gamma} \left( \frac{1}{2} I_t - Q_{t-1} \beta, 1 \right) \left( I_t - Q_{t-1} \beta \right)^{-1}, \frac{n(T-1)}{2} + q + 1 \right).
\end{align*}$$

Next, we find the full-conditionals for $I_t$:

$$\left[ I_t \mid \cdot \right] \propto \left[ y_t \mid I_t \right] \left[ I_{t+1} \mid I_t, \beta, \sigma^2 \right] \left[ I_t \mid I_{t-1}, \beta, \sigma^2 \right].$$

Here, it helps to rewrite the model as

$$I_t = M I_{t-1} + \nu_t,$$

$$\nu_t \sim \operatorname{Normal} \left( 0, \sigma^2 I \right),$$

where $M$ is a vector autoregressive propagator matrix, such that

$$M = \begin{cases} 
    x_i \beta_W + a_{ij} \beta_A, & \text{if } j = i \\
    b_{ij} (x_j - x_i)' \beta_B + a_{ij} \beta_A, & \text{if } j \in N_i \\
    a_{ij} \beta_A, & \text{if } j \not\in N_i.
\end{cases}$$

Also, to analytically find the full-conditional distribution for $I_t$, we can approximate the likelihood $\left[ y_t \mid I_t \right]$ using the empirical Bayes approximate likelihood.

$$\left[ y_t \mid I_t \right] \sim \operatorname{Normal} \left( \frac{700k}{5N}, 100k \hat{\theta}_t \left( 1 - \hat{\theta}_t \right) \right),$$

where,

$$\hat{\theta}_t = \frac{y_t}{100k},$$

$$k = 1,000.$$
Now we have the product of three Gaussian distributions for this full-condition $a_l$, which will result in another Gaussian distribution.

$$
\begin{align*}
[|I_t| \cdot] & \propto \exp \left\{ -\frac{1}{2} \left( y_t - I_t \frac{700k}{5N} \right)' \left( 100k \hat{\theta}_t (1 - \hat{\theta}_t) \right)^{-1} \left( y_t - I_t \frac{700k}{5N} \right) \right\} \\
& \times \exp \left\{ -\frac{1}{2} \left( I_{t+1} - MI_t \right)' \left( \sigma^2 I \right)^{-1} \left( I_{t+1} - MI_t \right) \right\} \\
& \times \exp \left\{ -\frac{1}{2} \left( I_t - MI_{t-1} \right)' \left( \sigma^2 I \right)^{-1} \left( I_t - MI_{t-1} \right) \right\}.
\end{align*}
$$

Thus,

$$|I_t| \cdot \sim \mathcal{N}(A^{-1}b, A^{-1}) ,$$

where,

$$A = \left( \frac{700k}{5N} \right) \left( 100k \hat{\theta}_t (1 - \hat{\theta}_t) \right)^{-1} \left( \frac{700k}{5N} \right) + M' (\sigma^2 I)^{-1} M + (\sigma^2 I)^{-1} ,$$

$$b = y' \left( 100k \hat{\theta}_t (1 - \hat{\theta}_t) \right)^{-1} \left( \frac{700k}{5N} \right) + I_{t+1} I_t M' (\sigma^2 I)^{-1} .$$

Similarly, the full-conditional for $|I_T|$ is given by

$$
[|I_T| \cdot] \propto \left[ y_T \left| I_T \right. \right] \left[ I_T \left| I_{T-1}, \beta, \sigma^2 \right. \right] .
$$

Again using the empirical Bayes approximate likelihood for $[y_T \left| I_T \right.]$, we have the product of two Gaussian distributions, which results in a Gaussian distribution for $|I_T| \cdot$.

$$
\begin{align*}
[|I_T| \cdot] & \propto \exp \left\{ -\frac{1}{2} \left( y_T - I_T \frac{700k}{5N} \right)' \left( 100k \hat{\theta}_T (1 - \hat{\theta}_T) \right)^{-1} \left( y_T - I_T \frac{700k}{5N} \right) \right\} \\
& \times \exp \left\{ -\frac{1}{2} \left( I_T - MI_{T-1} \right)' \left( \sigma^2 I \right)^{-1} \left( I_T - MI_{T-1} \right) \right\}.
\end{align*}
$$

Thus,

$$|I_T| \cdot \sim \mathcal{N}(A^{-1}b, A^{-1}) ,$$

where,

$$A = \left( \frac{700k}{5N} \right) \left( 100k \hat{\theta}_T (1 - \hat{\theta}_T) \right)^{-1} \left( \frac{700k}{5N} \right) + (\sigma^2 I)^{-1} ,$$

$$b = y' \left( 100k \hat{\theta}_T (1 - \hat{\theta}_T) \right)^{-1} \left( \frac{700k}{5N} \right) + I_{T-1} M' (\sigma^2 I)^{-1} .$$
Using these full-conditional distributions, we can implement a Gibbs sampler MCMC algorithm to draw random samples from each distribution to approximate the posterior distribution. We can then perform model selection using a variety of covariate sets to assess competing models for the GFT data.

3.4 Assessing Model Fit

Assessing model fit in the Bayesian framework is often done using the concept of deviance,

\[ D(y, \theta) = -2\log[p[y|\theta]], \]  

where, \( \theta \) generally represents the model parameter. In general, the model with the lowest expected deviance will have the highest posterior probability (Zhu and Carlin, 2000). Due to the fact that we have random parameters, we use:

\[ D_\theta(y) = D(y, \hat{\theta}(y)), \]  

where,

\[ \hat{\theta}(y) = E(\theta|y), \]

and

\[ D_{ave}(y) = E(D(y, \theta)|y) \approx \frac{1}{N} \sum_{k=1}^{N} D(y, \theta^{(k)}), \]  

then the effective number of parameters (i.e., unconstrained parameters) is

\[ p_D = D_{ave}(y) - D_\theta(y). \]  

In general, a parameter counts as 1 if it is estimated with no prior information, and as a 0 if all information comes from the prior distribution. \( p_D \) represents the decrease in deviance expected from estimating the parameters in the model (Zhu and Carlin, 2000). The expected predictive deviance can be used as a criterion of model fit, and is known as the deviance information criterion (DIC):

\[ DIC = D_{ave}(y) + p_D. \]  

Smaller DIC values, then, imply a better model fit (Zhu and Carlin, 2000).
3.5 Covariate Selection

Of the several factors that may have an effect on the transmission of influenza, we investigated the influence of air travel, population density, average temperatures, and the age of an individual. Each of these factors can be accounted for in the model through the covariate matrix $X$.

Air travel data were acquired from the "Origin and Destination Survey" database from the TranStats data library (http://www.transtats.bts.gov/). The number of passengers traveling from one state to another was accumulated for the year 2005, and stored in a matrix $A$ where $a_{ij}$ represents the number of passengers who traveled from state $i$ to state $j$. The matrix $A$ is then asymmetric, allowing for air travel between two states to be more concentrated in one direction than the other (e.g., more people may fly from Utah to New York than from New York to Utah).

Population density can be accounted for by dividing the Census Bureau state population estimates for 2008 by the state's respective area in square miles. To moderate the effect of larger states, we also account for area as a separate variable. This ensures that large states with high population densities (e.g., California) do not wrongfully dominate the potential effects of population density. Maps of population density and area are represented graphically in figures 3.1 a and b.

To account for change in temperature throughout the year, we use two variables: the minimum January temperature and the minimum July temperature for each state. By accounting for both winter and summer temperatures, we indirectly account for how moderate a state is. These variables are comprised of the minimum temperature each month over thirty years in various cities of each state. The city temperatures were averaged to obtain an overall minimum state temperature. Maps of the minimum January and July temperatures are shown in figures 3.1 c and d, respectively.

In this analysis, we are referring to influenza transmission between states rather than people directly, thus age will be represented as an age sensitivity variable for each state. The U.S. Census Bureau provided population estimates for 2005, separated by age category. To
create the age sensitivity variable, we sum the number of people whose ages are associated with increased risk (i.e., young and old) and divide by the number of people whose ages are not associated with increased risk. The “young” category was defined by ages ranging from 0 to 17 years; the “old” category was defined by 65 to 80+ years. Thus, a higher age sensitivity in a state represents a higher population of old and young people. The age covariate is illustrated as a map in figure 3.1 e.
Figure 3.1. Covariate maps: (a) population density (note that darker shades indicate higher population density), (b) state area (note that darker shades indicate larger area), (c) minimum January temperature (note that darker shades indicate warmer January temperatures), (d) minimum July temperature (note that darker shades indicate warmer July temperatures), and (e) age sensitivity (note that darker shades indicate increased age sensitivity).
CHAPTER 4
RESULTS

4.1 Model Selection

Several combinations of covariates were assessed in the model for this analysis. Each time a given model was fit, all covariates in each set appeared to be significant, likely because of the large sample size of the GFT dataset. In order to make inference on the coefficients, the first 1,000 MCMC samples (of 5,000) were discarded as a burn-in. Figure 4.1 suggests that this number should be sufficient. Table 4.1 provides a list of the covariates used and associated deviance statistics resulting from the fit for each of the models in the analysis. Each of these models also includes an air travel component.

<table>
<thead>
<tr>
<th>Set of Covariates</th>
<th>$D_\phi$</th>
<th>$D_{ave}$</th>
<th>$p_D$</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area, PPSM, Jan, July, Age, Jan×July,</td>
<td>124852.1</td>
<td>138828.1</td>
<td>13975.98</td>
<td>152804.1</td>
</tr>
<tr>
<td>Jan×Jan×Age, July×Age</td>
<td>124792.9</td>
<td>138779.3</td>
<td>13986.45</td>
<td>152765.8</td>
</tr>
<tr>
<td>Area, PPSM, Jan, July, Jan×July</td>
<td>124749.1</td>
<td>138733.9</td>
<td>13984.87</td>
<td>152718.8</td>
</tr>
<tr>
<td>PPSM, Jan, July, Jan×July</td>
<td>124729.8</td>
<td>138719.7</td>
<td>13989.96</td>
<td>152709.7</td>
</tr>
<tr>
<td>Area, PPSM, Jan, July, Age</td>
<td>124676.1</td>
<td>138681.6</td>
<td>14005.5</td>
<td>152687.1</td>
</tr>
<tr>
<td>Area, PPSM, Jan, July</td>
<td>124620.4</td>
<td>138624.5</td>
<td>14004.16</td>
<td>152628.7</td>
</tr>
</tbody>
</table>

Based on the DIC values in Table 4.1, the best model for these data consists of the covariates: area, population density, minimum January temperature, and minimum July temperature.

4.2 Model Results

Table 4.2 gives the posterior means and 95% credible intervals taken from the marginal posterior distributions for each of the covariate coefficients.

We assessed convergence visually and provide trace plots in figures 4.1 to illustrate the mixing of the chains.

The maps in figure 4.2 were created by taking the product $X E(\beta|y)$. The first map consists of $X E(\beta|y)$, thus creating a map of interstate transmission. Similarly, the second
Table 4.2. Posterior means (and 95% credible intervals) for covariate coefficient marginal posterior distributions

<table>
<thead>
<tr>
<th></th>
<th>Intrastate</th>
<th>Interstate</th>
<th>Air</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air</td>
<td>-</td>
<td>-</td>
<td>0.238 (0.234, 0.241)</td>
</tr>
<tr>
<td>Area</td>
<td>0.078 (0.071, 0.085)</td>
<td>0.103 (0.093, 0.113)</td>
<td>-</td>
</tr>
<tr>
<td>PPSM</td>
<td>0.165 (0.146, 0.185)</td>
<td>0.114 (0.096, 0.132)</td>
<td>-</td>
</tr>
<tr>
<td>January</td>
<td>-0.163 (-0.175, -0.151)</td>
<td>-0.168 (-0.194, -0.141)</td>
<td>-</td>
</tr>
<tr>
<td>July</td>
<td>0.142 (0.131, 0.153)</td>
<td>0.029 (0.001, 0.057)</td>
<td>-</td>
</tr>
</tbody>
</table>

The map consists of $X E(\beta_W|y)$, creating a map of intrastate transmission.

The map in figure 4.3 was created by calculating $\sum_{j=1}^{N_i} b_{ij} (x_j - x_i)^T \beta_B$ for each state of interest. This calculation represents ILI transmission into state $i$, thus creating a "susceptibility" surface.

Using the posterior mean of $I$ at each time, we can use equations 3.8 and 3.9 to estimate the number of people who are recovered and susceptible, respectively. Figure 4.4 displays these estimates for the state of Utah. These plots can be obtained for different states of interest. However, these calculations require large amounts of memory, so we provide Figure 4.4 to illustrate the additional forms of inference when using such models.
Figure 4.1. MCMC samples from the distributions of (a) the intrastate covariate coefficients (\(\beta_W\)), (b) the interstate covariate coefficients (\(\beta_B\)), (c) the air covariate coefficient (\(\beta_A\)), and (d) \(\sigma^2\). For (a-b), black represents the coefficient for area, red represents the coefficient for PPSM, green represents the coefficient for minimum January temperature, and blue represents the coefficient for minimum July temperature. In (d) the square root of the samples is displayed for simplicity.
Figure 4.2. Transmission surfaces: (a) flow surface of interstate transmission, where ILI transmits more easily between lighter states, and (b) intrastate transmission, where ILI increases more easily in darker states.

Figure 4.3. Map of susceptibility of state $i$ to ILI from state $j$, where darker states are more susceptible to obtaining ILI from surrounding states.
Figure 4.4. The Utah example of susceptible, infected, and recovered curves over time, where \( N \) represents the total population of Utah estimated using the 2000 census. (Note that the original GFT data occur at weekly intervals, but we are indexing the x-axis by month for clarity.)
Table 4.2 gives the posterior means and 95% credible intervals associated with each coefficient $\beta$ (i.e., $\beta_W$, $\beta_B$, and $\beta_A$). Here we see that each of the coefficients are significantly different from zero, as none of the credible intervals contain zero. Note that all effects are positive except that of the minimum January temperature for both intrastate and interstate. Thus, rather than discussing each effect separately, it may be more interesting to compare intrastate effects with the paired interstate effects. Here it is important to remember how the interstate covariates appear in the model. In equation 3.6, we see for state $i$, the difference between covariates in states $i$ and $j$, multiplied by a weight $b_{ij}$, and summed over the neighboring states of state $i$. Thus, a positive interstate coefficient, $\beta_B$, suggests that high covariate values in neighboring states of state $i$ yield an increase of transmission of ILI into state $i$.

In Table 4.2, we see that the air component $\beta_A$ has a positive effect. That is, states with higher rates of incoming air travel have an increase of ILI transmission. The area covariate suggests for both the intra- and interstate that a larger area has a positive relationship with ILI transmission. This effect is greater in the interstate case, suggesting that larger states transmit ILI to smaller states with greater ease than from smaller states into larger states. The population density (ppsm) covariate also has a positive relationship with ILI transmission. That is, states with higher population density will both transfer ILI to states with smaller population density and have a higher transmission rate within the state. Considering both temperature variables, we see that states which do not have moderate temperatures throughout the year (i.e., states that have very cold winters and very hot summers) have higher within state transmission rates. The interstate minimum January coefficient suggests that ILI is transmitted more readily from states with colder winters than states with warmer winters.

Figure 4.2 displays the transmission surfaces for interstate and intrastate transmission. Because $X\beta_B$ is a surface that, when differentiated in space, resembles the interstate trans
mission component of equation 3.6, \( \sum_{j=1}^{N_i} b_{ij} (x_j - x_i)' \beta_B \), darker areas indicate barriers to ILI transmission. We see, in Figure 4.2 (a), a barrier in the midwest states. This indicates a path that ILI takes each season, as ILI is more likely to spread through the lighter shaded states. In Figure 4.2 (b), darker areas indicate a higher rate of intrastate transmission. Here we see that states that are large or have a high population density have higher rates of transmission within themselves. Figure 4.3 indicates which states are most susceptible to transmission of ILI from surrounding states. Examining both the intrastate transmission map and the interstate susceptibility map, we see, for example, that Texas has high within-state transmission, but is not very susceptible to surrounding states. Then, from the interstate transmission map, we see that Texas is likely to transmit ILI to its neighboring states due to its large population and area.

Finally, we make inference on the SIR model. Figure 4.4 gives one such example for the state of Utah. We see that the majority of the population is susceptible to ILI at any given time. The line representing the number of infectious people has a seasonal pattern, using the log scale. We see that, on average, about \( e^{0.5} = 13,360 \) people are infectious with influenza-like illness year round. This is about 0.67% of the Utah population. During the peaks of the "infectious" line (i.e., winter seasons), there are between \( e^{10.5} = 36,315 \) and \( e^{11.5} = 98,715 \) people, between 1.83% and 4.87% of the total population, who are infectious with ILI. During the summer, only about \( e^{8.5} = 4,915 \), or 0.24% of people, are infectious with ILI. The number of recovered people has a similar seasonal pattern to the number of infectious people. On average, about \( e^{12} = 162,755 \) people, approximately 8.2% of the population, are recovering from influenza-like illnesses. This approximation is consistent with the idea that people have ILI between 0 and 5 times per year. Also, we note that, throughout the year, approximately 90% of the Utah population are susceptible to ILI.
CHAPTER 6
CONCLUSION

In this analysis, we were able to incorporate the epidemiological SIRS model into our statistical model to make scientific inference on the process of influenza. Using this framework, we were able to account for several covariates and perform model selection techniques to determine which set of covariates provided the best fit for our data. We found that the spread of ILI in the continental U.S. may be affected by covariates such as population density and minimum winter and summer temperatures. We have discovered which U.S. states are more susceptible to ILI, and what the major sources of ILI are for each state.

For future work, it may be interesting to account for influenza coming from other countries. This is likely to occur with Canada and Mexico, and could potentially occur through air travel between countries. To account for bordering countries, given data, we could simply create our neighborhood matrix of border states to include provinces of Canada or states of Mexico in the same way that U.S. states are listed. Air travel from other countries could be accounted for in a similar way to what is currently being done; that is, create an asymmetric matrix with origin and destination countries, along with the origin and destination states. Then, use that as the source for $a_{ij}$ in equation 3.6.

It also may be interesting to account for vaccination effects. This could be done easily by incorporating the approximate percentage of people who got vaccinated each year as a covariate. However, if data were available, it may be more interesting to include it in the SIRS model more directly. Since it is believed that vaccinations reduce risk of influenza for certain populations, we might set up the model in two pieces such that the vaccinated population has a separate SIRS model. We could then compare the two pieces to see if getting an influenza vaccination changes the number of infectious people.

The findings of this analysis could be used to slow the transmission of influenza-like illnesses between states by using more precautionary measures in states that are more susceptible to other states. Also, by analyzing the interstate transmission map in Figure 4.2, we can determine the general path that ILI will follow as it spreads. This can be used
to regulate travel restrictions during severe epidemics or pandemics in which the risks of
the disease are much higher than the cost of enforcing travel restrictions. Thus, by using
the results of this analysis and, perhaps, future work, we will be able to increase knowledge
of influenza characteristics in the U.S. and abroad.
REFERENCES


