Spatial and Spatio-temporal Epidemiology 1 (2010) 177–185

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste

Assessing North American influenza dynamics with a statistical SIRS model

Mevin B. Hooten [a,*], Jessica Anderson [a], Lance A. Waller [b]

[a] Department of Mathematics and Statistics, Utah State University, Logan, UT 84322, USA
[b] Department of Biostatistics, Emory University, Atlanta, GA 30322, USA

ARTICLE INFO

Keywords:
Agent-based model
Dynamical model
Epidemic
Flu
Hierarchical model
Spatio-temporal statistics

ABSTRACT

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 prepared on a weekly basis by Google Flu Trends. The Google Flu Trends team has provided a powerful new approach to collecting and reporting epidemiological data and, when used in conjunction with sophisticated statistical models, can allow for the identification and quantification of the flow of influenza across the continental United States. Our proposed methods, when conditioned on such a 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.

1. Introduction

1.1. Motivation

With the recent outbreaks of the H5N1 and H1N1 strains of the influenza virus, there is an increasing need for sophisticated methods of characterizing the spatio-temporal dynamics of these epidemic processes. New statistical models have been developed that allow researchers to intuitively represent dynamic behavior in epidemiological processes and are capable of identifying both barriers and corridors to the spread of disease over large spatial extents (e.g., Lawson and Zhou, 2005; Waller et al., 2007; Wheeler and Waller, 2008; Hooten and Wikle, in press). Additionally, advances in search engine and data mining technology have allowed for the collection and reporting of continental-scale epidemiological data at a high temporal resolution (Ginsberg et al., 2009).

The transmission of disease can be complicated at the small-scale (Altizer et al., 2006), however, important features of large-scale transmission may become apparent if a model can be specified that is able to partition the dynamics into scientifically meaningful and identifiable components that can be estimated using available data. Resulting statistical inference can then provide information about past, current, and future states and mechanisms of the epidemiological process.

More specifically, in what follows we present a general spatio-temporal modeling framework that is specified in terms of intuitive small-scale (i.e., fine scale) dynamics and allows us to learn about commonly studied, but unobserved, epidemiologic features of the population in addition to both pathways and barriers to the spread of influenza in the continental United States. Moreover, we are able to link the spread of influenza to various environmental and anthropogenic covariates in an effort to determine whether certain states function as epidemiological sources or sinks and how susceptible states are to both intrastate and interstate transmission of influenza-like illness (ILI).

1.2. Agent-based statistical models

Numerous modeling frameworks for statistically characterizing spreading phenomena have appeared in

* Corresponding author.
E-mail addresses: Mevin.Hooten@usu.edu (M.B. Hooten), jess.a@aggiemail.usu.edu (J. Anderson), lwaller@sph.emory.edu (L.A. Waller).

1877-5845/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.sste.2010.03.003

the literature in the past decade. Many of these adopt a hierarchical structure so that multiple sources of uncertainty can be accounted for explicitly in each of the model levels (Cressie et al., 2009). The Bayesian hierarchical modeling framework (Berliner, 1996) has proven to be a powerful method for solving large complex problems by breaking them up into smaller tractable conditional problems. Such models often involve three levels, commonly referred to as the data model, process model, and parameter model. Specifically, Wikle (2003) and Hooten and Wikle (2008) let the middle level (i.e., the process) be motivated by spatio-temporal partial differential equations, whereas, Hooten et al. (2007) use a discrete matrix model to characterize the dynamics and forecast the spread of invasive species on continental scales. Other approaches involving generalized linear models (for non-Gaussian data) and Markov random fields have also proven to be effective for describing dynamic spatio-temporal processes (Royle and Dorazio, 2008; Zhu et al., 2005).

Another form of model that has successfully been utilized for mimicking the spatio-temporal behavior of epidemics is sometimes referred to as an agent-based model (e.g., Grimm et al., 2005; Grimm and Railsback, 2005). These models are often called individual-based models when implemented in the literal sense, where an individual organism's behavior is modeled directly. In the broader context, agent-based models can be constructed with the focus on a reporting unit as an "agent," rather than an individual organism. In this sense, the models are scalable and could operate at the molecular level or, as in our case, at a much larger spatial level. Smith et al. (2002) implemented such a model, which is based on a simple set of rules governing fine (i.e., between agent) behavior that leads to complicated large-scale evolution, to describe the dynamic behavior of a rabies epidemic in raccoon populations in Connecticut. A full statistical implementation of this type of model is presented by Hooten and Wikle (in press) and allows for the estimation of important types of epidemic behavior such as: persistence, anisotropic and non-stationary transmission, and long-distance transmission. A critical feature of the models presented by Hooten and Wikle (in press) links transmission probability to the change in covariates, rather than to the covariates themselves. In this way, they are able to focus the movement of the disease on the dynamic nature of the heterogeneous cellular landscape rather than the static covariates themselves. That is, having modeled persistence (a cell-based epidemic process) in other model components, they rely on the change in landscape to identify neighborhood-based movement probabilities. A complication in the models of Hooten and Wikle (in press) is that they are working strictly with binary data and thus must use discrete probability distributions to properly accommodate the support of the response variable.

In the mathematical literature, the difference between "top-down" versus "bottom-up" dynamic model specifications is referred to as Lagrangian versus Eulerian (Turchin, 1998). An example of the Eulerian approach occurs in cases where the process level of a hierarchical model is formulated in terms of a partial differential equation (PDE, e.g., Wikle, 2003). In these settings, the mathematical form of the PDE dictates the overarching dynamic behavior inherent in the model. By contrast, the Lagrangian approach relies on rules governing the small-scale dynamics that are then scaled up to exhibit large-scale behavior. Turchin (1998) shows that certain Lagrangian models, when scaled up and approximated, take an Eulerian form; in these cases, the mathematical approximation implies that the Lagrangian models are more general. The statistical crux then, is in the estimation aspects when considering these models from an inverse perspective (i.e., learning about underlying mechanisms given observed data).

In other recent work, Wheeler and Waller (2008), use a combination of statistical methods to identify and include important features of a heterogeneous environment (such as barriers and pathways), into a hierarchical Bayesian model for characterizing epidemics. In their methods, Bayesian areal Wombling is employed to detect areas of rapid change in the underlying environment. This is similar in spirit to the use of a gradient landscape (i.e., directional derivatives of the covariate mean field) to connect movement of the epidemic to the underlying environment.

1.3. SIRS models

A common way to study epidemics at the population level is through compartment models such as the SIR model (i.e., susceptible → infected → recovered), where the entire population of interest can be partitioned into these three compartments at any given time. When immunity to an infectious disease is only temporary, the SIR model is extended to accommodate the ability of a recovered individual to become susceptible again. These are termed SIRS models, and, are often formulated as a coupled system of ordinary differential equations (ODE):

$$\frac{dS}{dt} = -aSI + cR$$
$$\frac{dI}{dt} = aSI - bI$$
$$\frac{dR}{dt} = bI - cR$$

where, the quantities S, I, and R represent the portion of the total population (N, where $N = S + I + R$) that is either susceptible, infected, or recovered, respectively. Note that this model depends on instantaneous quantities, but can be integrated and considered over intervals or periods of time. In fact, most measurements of S, I, or R are recorded as integrated quantities.

1.4. Influenza-like illness

The United States Centers for Disease Control and Prevention (CDC) reports spatio-temporal influenza data, on a weekly basis, in the percentage of total physician visits that are due to influenza-like illness (ILI). The actual description of ILI, according to the CDC, is a person with a fever of at least 100 degrees Fahrenheit and a cough or sore throat. Also, each person can average between 1 and 6 episodes of ILI per year. Clearly, not all ILI cases will actually be influenza, but record keeping is performed in this manner for convenience and efficiency. Recently, research-

ers at Google formally linked an influenza-related search query index with the CDC ILI and began reporting both current and historical modeled ILI from June in year 2003 – current as the number of patients $(y_{i,t})$ of out of 100,000 physician visits that have ILI for state i and week t (Ginsberg et al., 2009).

In their paper, Ginsberg et al. (2009) claim that by simply compiling and processing their search query data faster than the CDC, they are able to report ILI (http://www.google.org/flutrends/) 1–2 weeks ahead of the official *ILI* data. That is, they are not forecasting the ILI, only predicting it using search query data. Thus, $y_{i,t}$ is a noisy version of true $ILI_{i,t}$, but also more readily available. Accounting for this additional source of uncertainty is an important component of any further modeling and inference based on these data.

The weekly ILI data themselves can be most easily viewed as a set of time series on the same plot, where each time series represents $y_{i,t}/100,000$ (Fig. 1). The spatio-temporal dynamics of ILI can be illustrated best as a movie, however, we can also view the empirical orthogonal functions (EOFs) of ILI by decomposing the space-time series using a principal components approach and viewing the components (EOFs) as well as the resulting scores. In a space-time setting, the EOFs can be visualized as maps and the scores represent the corresponding time series indicating at which points in the temporal domain the spatial pattern in each EOF reoccurs (Preisendorfer, 1988). In this case, the first two EOFs account for approximately 90% of the overall spatio-temporal variability and are shown in Fig. 2. By contrasting the time series with the data in Fig. 1 we can see the scores from the first EOF correspond to the overall magnitude of ILI, where the scores from the second EOF indicate a contrast largely between southwestern and northeastern states. In this case, the peaks and troughs in the second set of scores provide an indication of which side of the continental United States the annual epidemic ILI wave begins on. For example, notice that the first wave starts in the west while the second one starts on the east.

These observed epidemic waves contain valuable information as to how ILI, as a spreading phenomenon, propagates within and between states. The model we present in the next section partitions this spreading behavior into multiple components that allow us to make conclusions about the components of ILI transmission in the continental United States.

2. Methods

In constructing a model for continental influenza, where state-level ILI data are available, we seek to incorporate SIRS dynamics based on state-level intra- and inter-state disease transmission. In doing so, we take a hierarchical approach where we can accommodate the discrete and integrated nature of the available data conditioned on unobserved SIRS quantities of interest. The dynamics relating these unobserved quantities are then specified in an intuitive fashion while using available scientific knowledge about the epidemiology of ILI.

2.1. Data model

Recall that Google Flu Trends reports state-level ILI data on a weekly basis in terms of $y_{i,t}$, a count out of 100,000 physician visits in state i during week t. Thus, we begin by specifying a stochastic model for the data. In this case, we consider the observed count $(y_{i,t})$ conditioned on an underlying ILI probability $(ILI_{i,t})$ to come from a binomial distribution: $y_{i,t} \sim \text{Binom}(100K, ILI_{i,t})$, where $K \equiv 1000$. Note that, for a state population of N_i, the number of peo-

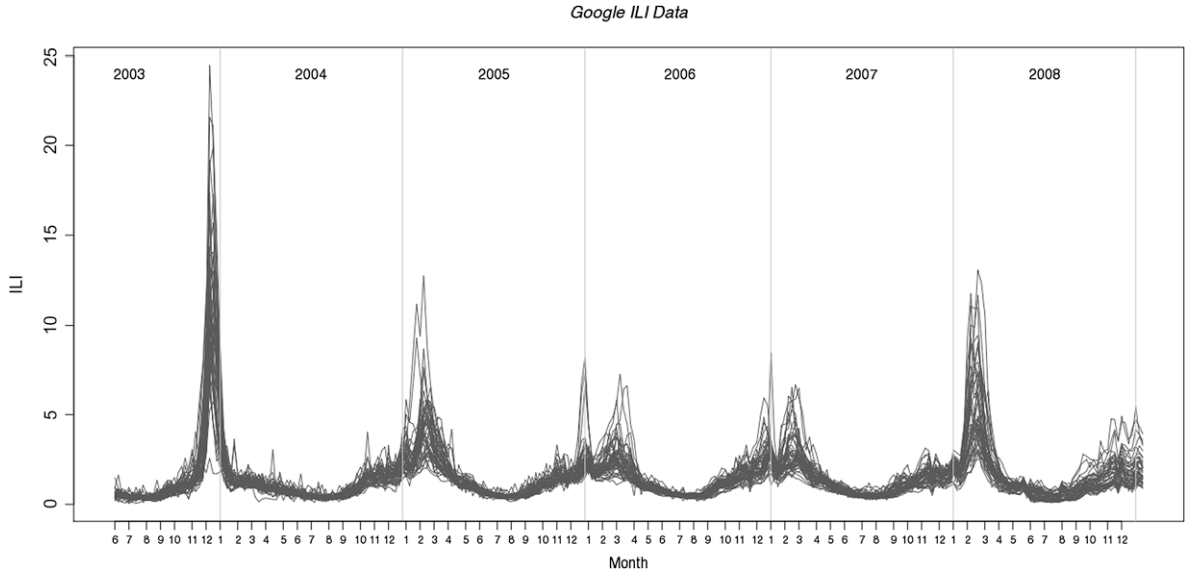

Fig. 1. The modeled ILI data provided by Google, represented here as a set of time-series where the year labels appear at the top of the figure and the x-axis labels represent the beginning of each corresponding month.

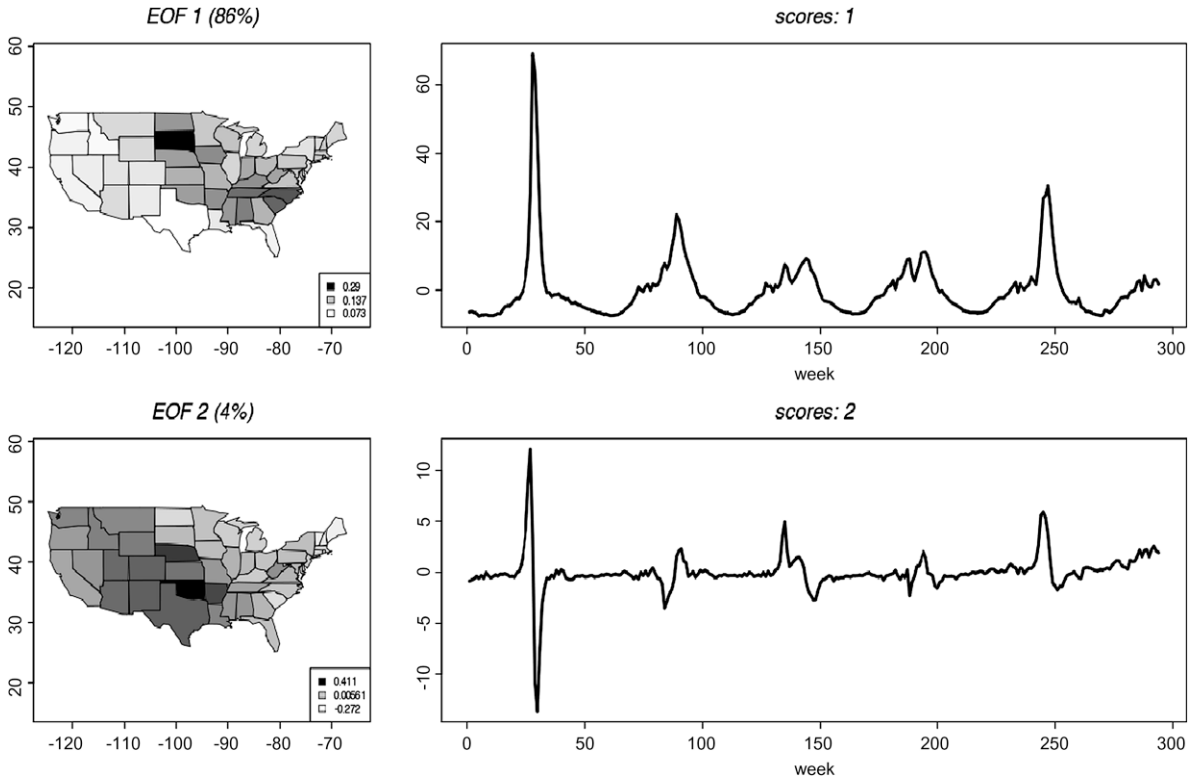

Fig. 2. The first two empirical orthogonal functions and associated scores corresponding to the ILI data over the continental United States.

ple who became infectious at time v in state i (i.e., $\widetilde{I}_i(v)$) is a function of $ILI_{i,t}$:

$$ILI_{i,t} = \frac{1}{N_i}\int_{t-1}^{t} \widetilde{I}_i(v)dv. \tag{1}$$

In order to connect this data model to an underlying SIRS model, we need to convert $\widetilde{I}_i(v)$ to the number of infectious people in state i at time t (i.e., $I_{i,t}$). This requires information about the persistence of infectiousness. We adopt a common model used for the natural history of influenza (e.g., Yang et al., 2009), where the duration of infectiousness extends from 1 to 7 days with probabilities $\mathbf{p} = (1, 1, 1, 0.8, 0.6, 0.4, 0.2)'$, for each of the days respectively. Assuming ILI infection occurs uniformly throughout the week, we arrive at the following relationship between $I_{i,t}$ and $\widetilde{I}_i(v)$:

$$I_{i,t} = \int_{t-1}^{t} p(v)\widetilde{I}_i(v)dv = \left(\frac{5}{7}\right)N_i ILI_{i,t}, \tag{2}$$

where the latter equality obtains from substituting (1) into (2). The result in (2) now implies the following data model: $y_{i,t} \sim \text{Binom}(100K, I_{i,t}(\frac{7}{5N_i}))$. Hence, assuming that N_i is known, we now have a likelihood such that the data depend on an unknown underlying epidemic process (i.e., $I_{i,t}$).

2.2. Process model

The underlying system dynamics are specified in the middle stage of the hierarchical model. In the specific case of ILI (i.e., a set of diseases) a process model based on compartment dynamics in which immunity is only temporary (i.e., SIRS) can be useful for studying transmission. Given the discrete temporal nature of the data, we use a coupled set of ordinary difference equations to represent the SIRS dynamics:

$$S_{i,t} = N_i - I_{i,t} - R_{i,t}, \tag{3}$$

$$I_{i,t} = \mathbf{x}_i'\boldsymbol{\beta}_W I_{i,t-1} + \sum_{j=1}^{n_i} b_{i,j}(\mathbf{x}_j - \mathbf{x}_i)'\boldsymbol{\beta}_B I_{j,t-1} + \sum_{j=1}^{n} a_{i,j}\beta_A I_{j,t-1} + \eta_{i,t}, \tag{4}$$

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

where, $\eta_{i,t} \sim \text{N}(0, \sigma^2)$ and the equation for the susceptible portion (3) of the population is self-explanatory (essentially, every person that is not either infected or recovered in state i and time t), while the other two equations, (4) and (5), are discussed in detail in the following sections.

2.2.1. Infected component

Note that, in (4), the model for infection contains three main additive terms. The first, $\mathbf{x}_i'\boldsymbol{\beta}_W I_{i,t-1}$, corresponds to intrastate transmission and is a function of state-level covariates $\mathbf{x}_i$. As a dynamic model, this term expresses the number of infected people at the current time as a function of the number of infected people at the previous time and unique state-level characteristics. The $\boldsymbol{\beta}_W$ coeffi-

cients, then, represent the importance of each covariate to intrastate transmission. The second term, $\sum_{j=1}^{n_i} b_{i,j}(\mathbf{x}_j - \mathbf{x}_i)'\boldsymbol{\beta}_B I_{j,t-1}$, is the first of two terms dealing with interstate transmission. In this case, it is a function of the differential between state characteristics and interacts with all contiguous states (of which there are n_i) infection status at the previous time. The differences, $\mathbf{x}_j - \mathbf{x}_i$, are used here to describe an underlying flow surface, similar to that considered by Hooten and Wikle (in press), in terms of directional derivatives. The $b_{i,j}$ are weights that represent the connectivity of states i and j; in this case they represent the proportion of shared boundary. The coefficients, $\boldsymbol{\beta}_B$ are most easily visualized when they are premultiplied by the covariate matrix $\mathbf{X} \equiv (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$. The quantity $\mathbf{X}\boldsymbol{\beta}_B$ represents the flow surface itself which can be thought of as a wavy map where ILI travels easily from high values to low values. This surface, when estimated, provides one of the main forms of inference allowed by the model. It should also be noted that this state to state transmission is easily specified and the main reason why we describe this model as agent-based.

The final non-error term in (4), $\sum_{j=1}^{n} a_{i,j}\beta_A I_{j,t-1}$, accommodates interstate transmission through air travel, a second form of connectivity between states. In this equation, the $a_{i,j}$ represent the relative amount of air travel from state j to state i and the coefficient β_A then scales this as necessary to provide the contribution of long-distance transmission during period $(t-1, t]$.

2.2.2. Recovered component

Noting that the recovered portion of the population is known when the $I_{i,t}$ are known, we simply need to formally define the relationship between $R_{i,t}$ and $I_{i,t}$. In doing so, we first rewrite (5) as:

$$R_{i,t} = R_{i,t-1} + I_{i,t}\left(\frac{2}{5}\right) - cR_{i,t-1} = R_{i,t-1}(1-c) + I_{i,t}\left(\frac{2}{5}\right) = \sum_{j=1}^{R_{i,t-1}} \left(1 - \frac{u_j}{52}\right) + I_{i,t}\left(\frac{2}{5}\right), \quad (6)$$

where, traditionally, the coefficient c controls the amount of recovered people who become susceptible in week t. In the latter equality, we express this temporary immunity in terms of the random variables u_j, for the jth recovered person at time t. Due to the previously described natural history of ILI, a person who has had an ILI can become infected with another ILI u_j more times in a year, where $u_j \sim \text{discunif}(0, 5)$.

For the final term in (6), we note that, as a consequence of our specification in (2), each week, a total of $ILI_{i,t}N_i\left(\frac{2}{7}\right)$ people recover. Thus, since $ILI_{i,t} = \frac{7I_{i,t}}{5N_i}$, we are left with $I_{i,t}\left(\frac{2}{5}\right)$ people that were infected but are now recovered.

2.3. Parameter model

A Bayesian approach allows us to easily accommodate the multiple sources of stochasticity in the hierarchical model specified in the previous sections. Thus, in completing the full hierarchical specification we need to specify probability distributions for the unknown random parameters of interest. Given that many of the model parameters are assumed to be fixed and known (based on descriptions in the literature), we have only to specify distributions for the interstate transmission coefficients ($\boldsymbol{\beta}_W, \boldsymbol{\beta}_B$, and β_A) and the variance component (σ^2). Specifically, we used a $N(\mathbf{0}, \boldsymbol{\Sigma}_\beta)$ prior for $\boldsymbol{\beta}$ and a uniform on $(0, \gamma)$ for the standard deviation σ (as suggested by Gelman, 2006). Note that, in the prior specification above, $\boldsymbol{\beta} = (\boldsymbol{\beta}_W', \boldsymbol{\beta}_B', \beta_A')'$ and $\boldsymbol{\Sigma}_\beta = 1000 \cdot \mathbf{I}$ while γ is chosen to be a very large finite value such that the prior on σ is proper, but has minimal effect on the posterior.

2.4. Implementation

In a Bayesian framework, we are interested in the posterior distribution of the unobserved latent process S, I, R, as well as the unknown random parameters $\boldsymbol{\beta}$ and σ^2. Using the conventional square bracket notation for hierarchical Bayesian models, the posterior distribution can be written:

$$[\{\mathbf{S}_t\}, \{\mathbf{I}_t\}, \{\mathbf{R}_t\}, \boldsymbol{\beta}, \sigma^2 | \{\mathbf{y}_t\}] \propto \prod_{t=2}^{T}[\mathbf{y}_t|\mathbf{I}_t] \prod_{t=2}^{T}[\mathbf{I}_t|\mathbf{I}_{t-1}, \boldsymbol{\beta}, \sigma^2][\boldsymbol{\beta}][\sigma^2]. \quad (7)$$

The posterior distribution in (7) can be easily approximated using Markov chain Monte Carlo (MCMC) with Metropolis-Hastings updates for non-conjugate parameters (Gelman et al., 2004). For computational efficiency, we take an additional approximation step in the implementation, using a normal approximation to the binomial likelihood. Given the actual observed values of $y_{i,t}$ we have found this to be a very reasonable approximation that yields conjugate full-conditional distributions for all unknown quantities and allows for the use of a Gibbs sampler MCMC algorithm.

3. Results

For covariates, we consider four spatial variables that we believe may be influential for the spread of ILI in the United States (Fig. 3):

- Size of the state (area in square miles).
- Population density in the state (population per square mile).
- Winter temperature (mean January temperature).
- Summer temperature (mean July temperature).

Using weekly Google Flu Trend data from June 1, 2003 to January 4, 2009, along with the covariates depicted in Fig. 3 and a large sample of airline ticket data (http://www.transtats.bts.gov/), we fit the model discussed in the previous section to obtain parameter and flow surface estimates. The Gibbs sampler, which was written in the R Statistical Programming Environment (R Development Core Team, 2008), was run for 5000 iterations (taking approximately 5 h on an 8 processor 3 Ghz server with 28 GB of RAM). Using 3 MCMC chains, the Gelman–Rubin convergence statistics (Gelman and Rubin, 1992) were computed for each parameter and were all less than

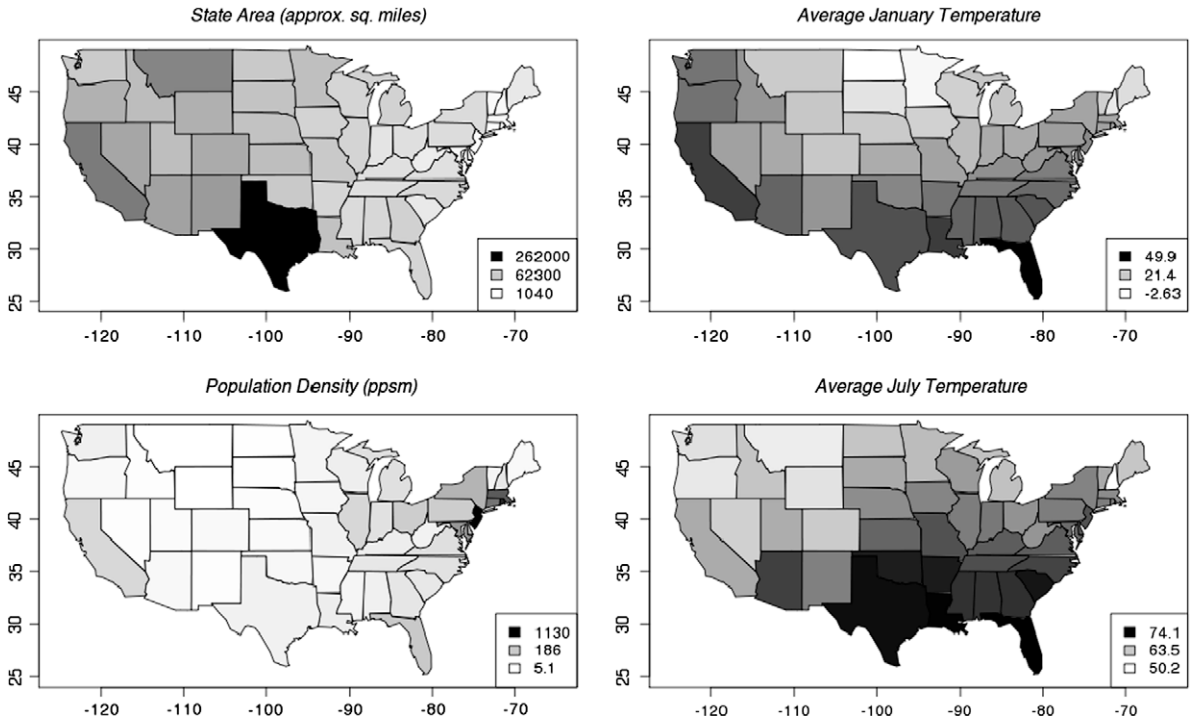

Fig. 3. State-level covariates used in the SIRS process model.

1.001, indicating acceptable MCMC convergence was achieved. For inference, the first 1000 MCMC samples were discarded as burn-in, and the remainder were used to approximate the transmission coefficients $\boldsymbol{\beta}$, which are provided in Table 1. Equal tail posterior credible intervals indicate that all of the coefficients are significantly different from zero (at the $\alpha = 0.05$ level). The estimated posterior mean spatial fields ($\mathbf{X}\boldsymbol{\beta}_W$ and $\mathbf{X}\boldsymbol{\beta}_B$) are shown in Fig. 4.

Another means for illustrating the results are to view the individual state posterior susceptibilities, both in terms of interstate (Fig. 5) and intrastate susceptibility (Fig. 6). We refer to intrastate susceptibility as $\mathbf{X}\boldsymbol{\beta}_W$, while interstate susceptibility corresponds to the cumulative effect of all neighboring states on the state in question and is computed as: $\sum_{j=1}^{n_i} b_{i,j}(\mathbf{x}_j - \mathbf{x}_i)'\boldsymbol{\beta}_B$, where the sum is over all neighbors of state i.

In addition to the full model discussed above, we also fit two additional reduced models: one omitting the between-state transmission (i.e., middle term on the right hand side of (4)), and another omitting the air travel component of transmission (i.e., second to last term in (4)). The deviance information criterion (DIC, Gelman et al., 2004) resulting from each of these model fits were 152,615, 152,569, and 152,315, respectively.

4. Discussion

The results from fitting the full model suggest that all model parameters are significant in the model, and thus, deemed important for describing ILI dynamics in the continental United States. For example, the first column of Table 1 indicates that state area, population density (PPSM), and average July temperature are all positively related to intrastate influenza transmission, while average January temperature had a negative effect. These results imply that as area, population, and summer temperature increase within a state, so do the transmission rates; whereas, as winter temperatures get colder, interstate transmission increases. We find the area, population, and winter temperature results to have reasonable scientific interpretations,

Table 1
Posterior means (and 95% credible intervals) for model coefficients ($\boldsymbol{\beta}_W, \boldsymbol{\beta}_B, \beta_A$), arranged in terms of their respective covariates. The covariates were all standardized by subtracting the mean and dividing by the standard deviation for the purposes of comparing coefficient values. The resulting 95% posterior credible intervals for the coefficients suggest all parameters are significantly different from zero.

	Intra	Inter	Air
Air	–	–	0.238 (0.23, 0.24)
Area	0.078 (0.072, 0.085)	0.103 (0.093, 0.113)	–
PPSM	0.166 (0.147, 0.185)	0.114 (0.096, 0.132)	–
January	−0.163 (−0.176, −0.151)	−0.168 (−0.194, −0.139)	–
July	0.142 (0.131, 0.153)	0.029 (0.002, 0.055)	–

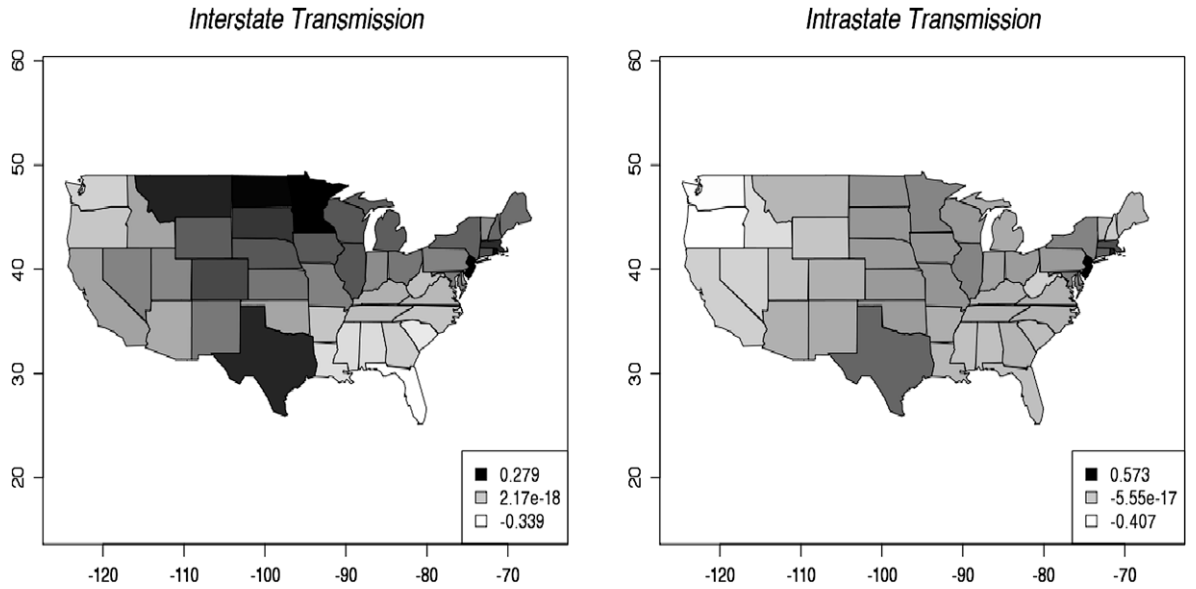

Fig. 4. Map 1 (left) represents the posterior mean flow surface ($\mathbf{X}\boldsymbol{\beta}_B$) over which the least resistance of spread is from high to low (i.e., dark to light) values. Map 2 (right) represents the susceptibility of each state to intrastate transmission of ILI (i.e., $\mathbf{X}\boldsymbol{\beta}_W$).

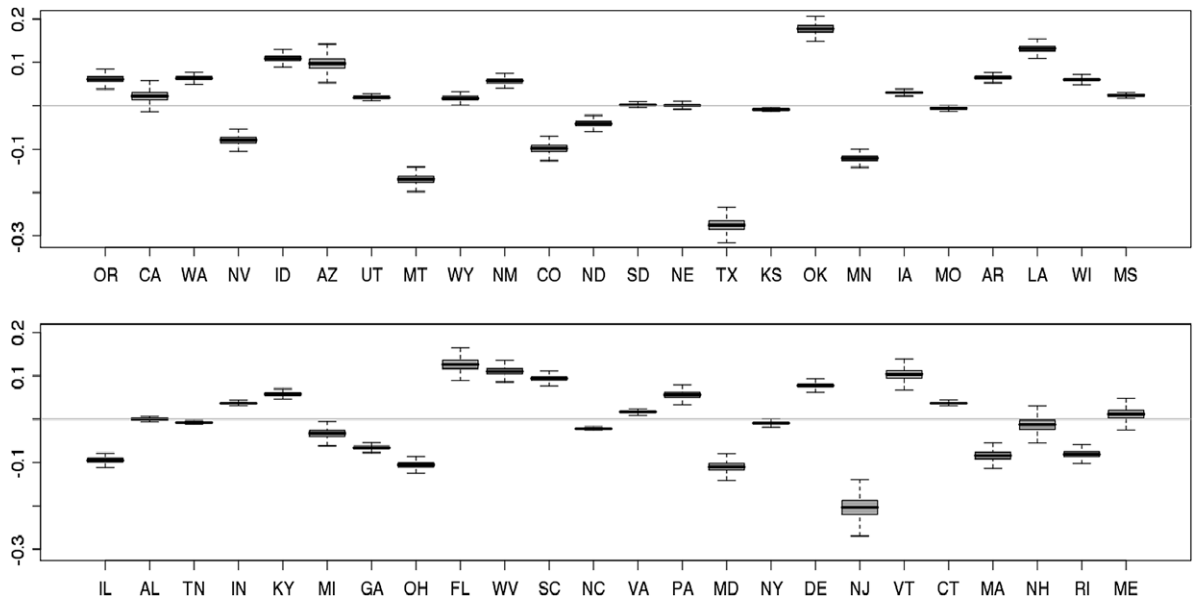

Fig. 5. Interstate susceptibility displayed as a sequence of boxplots, for all states, arranged from west to east.

while the summer temperature results require more thought; it is not immediately apparent why an increase in July temperature should lead to increased intrastate ILI transmission. Of course, given the observational nature of this study, it is quite likely that estimated effects are confounded with lurking variables not included in the study. However, regardless of potential multicollinearity between any missing covariates and those considered here, inference concerning the underlying flow surfaces can still be made. For example, Map 2 in Fig. 4 indicates (with darker shades of gray) which states are more sensitive to intrastate transmission than others. This map implies that many of the states with large areas in the middle of the country are more susceptible to intrastate transmission than those in, say, in the northwest (e.g., Oregon and Washington) where January temperatures are higher than other states at similar latitudes. State-level transmission could also be viewed as susceptibility, and Fig. 6 identifies which of the continental states are most and least susceptible to intrastate transmission. The boxplots shown in Fig. 6 represent the marginal posterior distributions of $\mathbf{X}\boldsymbol{\beta}_W$ given the Google Flu Trends data and available covar-

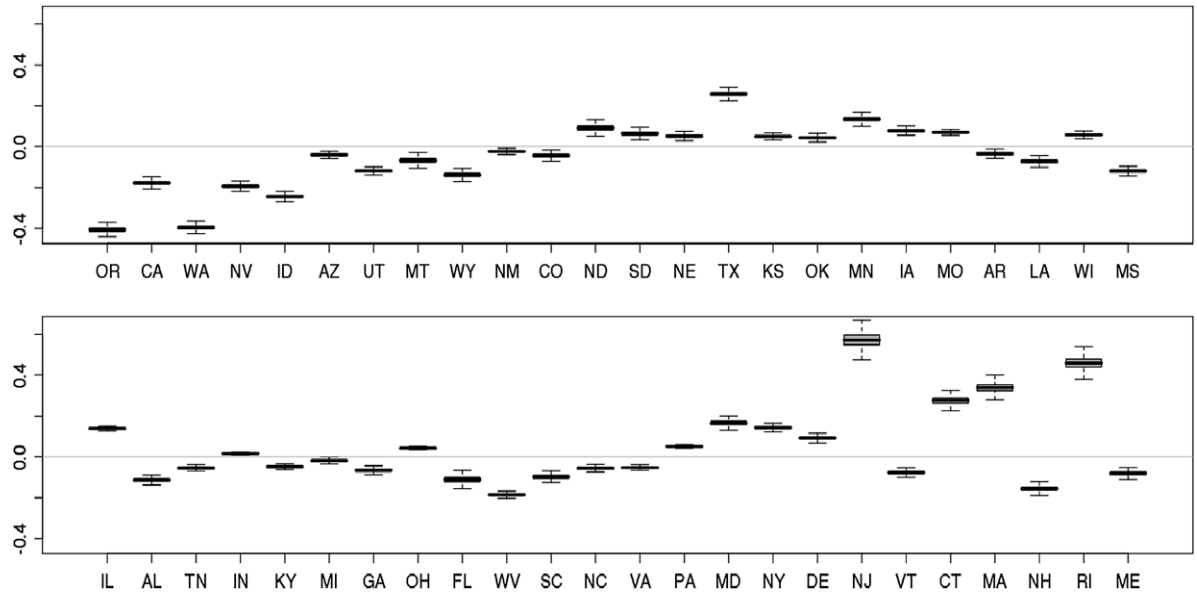

Fig. 6. Intrastate susceptibility displayed as a sequence of boxplots, for all states, arranged from west to east.

iates. Thus, boxplots that are significantly larger than zero indicate the most susceptible states, while those near zero indicate a neutral intrastate susceptibility. States with negative posterior susceptibilities (e.g., Oregon and Washington) suggest that ILI transmission is likely due to other sources (e.g., interstate transmission).

In terms of interstate transmission, we can see from the second column of Table 1 that the effects of the covariates on the interstate flow surface share the same sign, but differ in magnitude, from those influencing intrastate transmission. From Table 1 it is apparent that all coefficients except average January temperature have positive posterior mean effects, though the role that average July temperature plays in interstate transmission is reduced. These results imply that if a neighboring state is larger, has a higher population density, and higher summer temperature than the state of interest, the interstate transmission increases; the converse is true for neighboring states with higher winter temperatures, implying that ILI spreads slower from states with warm winters to states with cold winters. Viewed as a map, the posterior mean interstate flow surface (Map 1, Fig. 4), corresponding to $\mathbf{X}\boldsymbol{\beta}_B$, illustrates where these areas of high and low interstate transmission occur. The darker northern Great Plains states and midwest indicate a ridge on the surface. Theoretically, this flow surface would suggest that if an ILI wave began in the northern Great Plains, it would disperse rapidly "downhill" toward the southeastern states and west coast. We do not see this behavior in the data however, and thus it is more likely that the ridge of higher values in the center of the country represents a distinct barrier to cross-country transmission. That is, an epidemic starting on the west coast will spread readily through western states, while only slowly making its way through the center of the country toward the east. When considering this finding in combination with that of Map 2 in Fig. 4, it becomes obvious that interstate ILI transmission in the middle of the country actually occurs relatively slowly and thus the intrastate mechanism becomes more of a driver for the epidemiologic process in that region. Generally, darker contiguous regions on such maps indicate barriers to disease spread, while lighter contiguous regions suggest potential corridors for spread. Overall, the slightly larger DIC values for the full versus reduced model (where interstate transmission is omitted) suggest that the component of the infection model (4) involving interstate transmission may not be contributing much to the overall fit of the model. In computing the DIC, we found that the effective number of parameters were about four greater in the full model than the reduced model without the between state component. In this case, if one seeks inference regarding interstate transmission, it may be beneficial to account for these effects even though the DIC is slightly larger when they are included.

The interstate susceptibility can also be viewed on a state by state basis (Fig. 5). The boxplots in Fig. 5 represent the approximate posterior distribution for the sum of incoming ILI transmission from neighboring states. That is, states with posterior susceptibilities greater than zero indicate that neighboring states are responsible for a significant portion of ILI transmission. Based on the available data and covariate information used in this study, Figs. 5 and 6 indicate that small states with high population densities, such as New Jersey and Rhode Island, are most susceptible to intrastate ILI transmission, while states surrounded by larger states with high population densities act as sinks for ILI and are quite susceptible to neighborhood-based interstate transmission. In terms of long-distance interstate transmission, the positive posterior mean coefficient for air travel (Table 1) indicates that air travel may play a significant role in the spread of ILI, beyond the intrastate and neighborhood-based interstate trans-

mission. However, in this case, the resulting DIC was lower for the reduced model omitting the air travel component than either of the other two fitted models. Thus, despite a non-zero estimated air travel effect when fitting the full model, the air travel component may not be helpful overall, as including it increases the number of effective parameters by approximately 50. In terms of modeling the spread of human influenza, this could be welcome news, as airline information can be difficult to obtain.

In summary, we have introduced a general hierarchical framework for utilizing the ILI data produced by Google Flu Trends to link the dynamics and mechanisms (including both barriers and pathways) of influenza dispersal in the continental United States. In this effort, we have built on the previous works of Wheeler and Waller (2008) and Hooten and Wikle (in press) and are able to avoid some of the complications arising from those earlier studies due to the support and extent of the data. The general agent-based framework we present here relies only specification of small-scale (within and between state) dynamics and can readily be extended to accommodate important features of continental epidemic spread such as time-varying covariates, different disease natural histories, continental boundary conditions (e.g., other countries), and other forms of data. Our model formulation focuses on separating intrastate dynamics from interstate dynamics using a relevant set of environmental and anthropogenic covariates in both static and gradient forms. Our model results provide an indication that valuable scientific learning can result when these types of models are combined with data arising from sophisticated search engine mining technology.

Acknowledgements

The authors would like to thank Rajan Patel and Google for advice and support for this project.

References

Altizer S, Dobson A, Hosseini P, Hudson P, Pascual M, Rohani P. Seasonality and the dynamics of infectious diseases. Ecol Lett 2006;9:467–84.

Berliner LM. Hierarchical Bayesian time-series models. In: Maximum entropy and Bayesian methods. The Netherlands: Kluwer Academic Publishers; 1996. p. 15–22.

Cressie NAC, Calder CA, Clark JS, Ver Hoef JM, Wikle CK. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecol Appl 2009;19:553–70.

Gelman A. Prior distributions for variance parameters in hierarchical models. Bayesian Anal 2006;1:515–33.

Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. 2nd ed. Boca Raton: Chapman & Hall/CRC; 2004.

Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Stat Sci 1992;7:457–511.

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinksi MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457:1012–4.

Grimm V, Revilla E, Berger U, Jeltsch F, Mooij WM, Railsback SF, Thulke H-H, Weiner J, Wiegand T, DeAngelis DL. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. Science 2005;310:987–91.

Grimm V, Railsback SF. Individual-based modeling and ecology. Princeton, NJ: Princeton University Press; 2005.

Hooten MB, Wikle CK. A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian collared-dove. Environ Ecol Stat 2008;15:59–70.

Hooten MB, Wikle CK. Statistical agent-based models for discrete spatio-temporal systems. J Am Stat Assoc, in press.

Hooten MB, Wikle CK, Dorazio RM, Royle JA. Hierarchical spatio-temporal matrix models for characterizing invasions. Biometrics 2007;63:558–67.

Lawson AB, Zhou H. Spatial statistical modeling of disease outbreaks with particular reference to the UK foot and mouth disease (FMD) epidemic of 2001. Prev Vet Med 2005;71:141–56.

Preisendorfer RW. Principal component analysis in meteorology and oceanography. Elsevier; 1988.

R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2008. ISBN 3-900051-07-0. URL http://www.R-project.org.

Royle JA, Dorazio RM. Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations, and communities. Academic Press; 2008.

Smith DL, Lucey B, Waller LA, Childs JE, Real LA. Predicting the spatial dynamics of rabies epidemics on heterogeneous landscapes. Proc Natl Acad Sci USA 2002;99:3668–72.

Turchin P. Quantitative analysis of movement. Sinauer Associates, Inc. Publishers; 1998.

Waller LA, Goodwin BJ, Wilson ML, Ostfeld RS, Marshall SL, Hayes EB. Spatio-temporal patterns in county-level incidence and reporting of Lyme disease in the northeastern United States, 1990–2000. Environ Ecol Stat 2007;14:83–100.

Wheeler DC, Waller LA. Mountains, valleys, and rivers: the transmission of raccoon rabies over a heterogeneous landscape. J Agric Biol Environ Stat 2008;13:388–406.

Wikle CK. Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology 2003;84:1382–94.

Yang Y, Sugimoto JD, Halloran ME, Basta NE, Chao DL, Matrajt L, Potter G, Kenah E, Longini Jr IM. The transmissibility and control of pandemic influenza A (H1N1) virus. Science 2009;326:729–33.

Zhu J, Huang H-C, Wu C-T. Modeling spatial-temporal binary data using markov random fields. J Agric Biol Environ Stat 2005;10:212–25.