Biometrics 000, 1–23 DOI: 000
000 2006

An Empirically Adjusted Approach to Reproductive Number Estimation for Stochastic Compartmental Models: A Case Study of Two Ebola Outbreaks

Grant D. Brown[1,*], Jacob J. Oleson[1], and Aaron T. Porter[2]

[1]Department of Biostatistics, University of Iowa, Iowa City, Iowa, U.S.A.

[2]Department of Applied Mathematics and Statistics, Colorado School of Mines, Golden, Colorado, U.S.A.

email: grant-brown@uiowa.edu

Summary: The various thresholding quantities grouped under the 'Basic Reproductive Number' umbrella are often confused, but represent distinct approaches to estimating epidemic spread potential, and address different modeling needs. Here we contrast several common reproduction measures applied to stochastic compartmental models, and introduce a new quantity dubbed the 'empirically adjusted reproductive number' with several advantages. These include: more complete use of the underlying compartmental dynamics than common alternatives, use as a potential diagnostic tool to detect the presence and causes of intensity process underfitting, and the ability to provide timely feedback on disease spread. Conceptual connections between traditional reproduction measures and our approach are explored, and the behavior of our method is examined under simulation. Two illustrative examples are developed: First, the single location applications of our method are established using data from the 1995 Ebola outbreak in the Democratic Republic of the Congo and a traditional stochastic SEIR model. Second, a spatial formulation of this technique is explored in the context of the ongoing Ebola outbreak in West Africa with particular emphasis on potential use in model selection, diagnosis, and the resulting applications to estimation and prediction. Both analyses are placed in the context of a newly developed spatial analogue of the traditional SEIR modeling approach.

Key words: Spatial SEIR; Spatial epidemiology; Model selection; Underspecification; Disease modeling; Epidemic prediction.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

The basic reproductive number, $\mathcal{R}_0$, is an important quantity in epidemiology. While its interpretation must be adapted to the problem of interest, in general terms, the basic reproductive number captures the expected number of secondary infections produced by a single infected individual in an entirely susceptible population. This seemingly intuitive definition is complicated by authors' varying implementations, which generally share the same thresholding properties but carry different interpretations. This is especially confusing given the similar terminology used in stochastic and deterministic epidemic models. Heffernan et al. (2005) note that: *"Surveying the recent literature, it quickly becomes apparent that a number of related quantities, all of which share this 'threshold' behavior, are used as surrogates for* $\mathcal{R}_0$*. For example,* $\mathcal{R}_0^n$ $(n > 0)$ *will give an equivalent threshold, but does not give the number of secondary infections produced by a single infectious individual."* The authors go on to describe several commonly used approaches to $\mathcal{R}_0$ estimation in deterministic models which may or may not give accurate estimates of the traditionally defined basic reproductive number, including examining stability conditions, testing for the existence of a disease free equilibrium, and characteristic equation methods.

Other authors define several reproductive numbers. In particular, a traditional estimate of $\mathcal{R}_0$ is sometimes presented alongside a temporally varying generalization, and a scaled version sometimes termed the "effective reproductive number" (Lekone and Finkenstädt, 2006; Chowell et al., 2004). The latter measure, discussed below, departs from the traditional definition of $\mathcal{R}_0$, and yet incorporates only a portion of the disease dynamics described by the encompassing modeling framework.

In part, this diversity may be attributed to the fact that the basic reproductive number, as usually defined, is somewhat removed from actual epidemics. The quantity requires that the modeler posit an entirely susceptible population, and takes epidemic behavior at a singular

instant (or window for discrete models) in time, mathematically extrapolating said behavior into the future. In contrast, real epidemics can be extremely variable, and are not always well summarized by a single number. Here we examine a different approach to estimating the reproductive characteristics of epidemics, and demonstrate improved ability to detect changes in transmission behavior not explicitly accounted for by the model. Termed the 'empirically adjusted reproductive number' and denoted $\mathcal{R}^{(EA)}$, we define a quantity which is easily derived for any discrete time stochastic epidemic model. It is developed here for a flexible class of spatial epidemic models known as stochastic spatial SEIR models.

Simply put, we are interested in estimating the expected number of secondary infections which a particular individual from a particular spatial location will cause in real populations: a measure uniquely suited to detecting changes in epidemic behavior for underspecified models and providing early indications of the effectiveness of intervention efforts. Hethcote (2000) defines a similarly motivated quantity known as the replacement number for deterministic versions of the models discussed here. However, simple intuitions from deterministic epidemic models do not always translate to stochastic formulations; in particular, the replacement number is defined to be strictly less than the basic reproductive number, while the true number of secondary infections caused per infectious individual can vary widely over the course of real and stochastically simulated epidemics.

After a brief review of current compartmental modeling techniques, we introduce the stochastic spatial SEIR model class, following many standard conventions for the specification of hierarchical models (Cressie and Wikle, 2011). While the general approach to $\mathcal{R}^{(EA)}$ estimation described here is easily applicable to a wide array of epidemic models, this family provides a natural and flexible framework for its development. In this setting, we compare several common reproductive numbers and derive our own, considering also the conceptual and mathematical relationships between them. Finally, the practical implications

of reproductive number choice are explored via several simulations and two examples: the 1995 Ebola outbreak in the Congolese city of Kikwit, and the 2014 Ebola epidemic in West Africa. The first example illustrates the strong contrast, even in well studied epidemics, between reproductive numbers for both well and underspecified intensity processes. The second demonstrates the applicability of this technique to emerging and evolving epidemics in real time over multiple spatial locations, and emphasizes the oversimplification of disease dynamics which can occur with traditional epidemic thresholding parameters.

2. The Spatial SEIR Model

2.1 *Background*

Compartmental models have a long history in the epidemic modeling literature, beginning with the SIR technique specified by Kermack and McKendrick (1927). These models are named for the disease states, or compartments, which define them. The most commonly used disease states are **S**usceptible, **E**xposed, **I**nfectious, and **R**emoved. These categories describe individuals in a population who are, respectively, able to contract an infection, are infected but not yet infectious, have become capable of spreading an infection, and are permanently removed from the susceptible population by death or recovery with immunity.

Numerous extensions to this framework have been developed in the intervening years, though only recently have stochastic formulations received a thorough spatial treatment. Cook et al. (2007) introduce heterogeneity in two ways: by assuming the existence of favorable and unfavorable hosts, and by imposing a network based contact structure in what they call an S-I percolation model. Another approach was developed by Hooten et al. (2011), who used an additive approach to estimating infectivity within and between contiguous spatial units for an SIRS model of influenza data. Taking a different perspective, Deardon et al. (2010) introduce individual level models incorporating measures of distance into contact

probabilities between epidemiological units. Much of this generalization has been done in the context of the recent foot and mouth disease outbreaks in the U.K., developing heterogeneous, spatially distributed, and partially censored epidemic processes (Jewell et al., 2009; Chis Ster et al., 2009; Chis Ster and Ferguson, 2007).

Additional work has focused on household models, where houses define population clusters and SIR or SEIR compartment structures provide a model for disease transmission (e.g. Cauchemez et al. (2004, 2009); van Boven et al. (2010)). Such techniques generally consider epidemic spread within population clusters and with exogenous sources, whereas our approach focuses on pathogen spread within and between population clusters. Compartmental dynamics have also been examined over analogous social or contact networks, most commonly from a simulation perspective (Verdasca et al., 2005; Keeling, 2004).

Sattenspiel and Dietz (1995) model spatial heterogeneity in a SIR framework by incorporating contact probabilities between spatial locations weighted by travel propensity and return rates in a mover-stayer framework, while Jandarov et al. (2014) consider epidemic spread to decrease with distance between units in an approximation to a gravity model. As spatiotemporal epidemic models present considerable computational challenges, many authors pursue hybrid approaches, which combine stochastic simulation and systems of ordinary and partial differential equations (LaBute et al., 2014). The fully probabilistic work of Porter and Oleson (2014) most closely matches our spatial development.

2.2 *Data Model*

A first step in building hierarchical epidemic models is to establish the relationship between the observed data and the model parameters; while the most common approach to this problem is to assume that a particular quantity is accurately and completely observed, it is often more reasonable to assume that some other relationship exists. For a spatial SEIR model over time points $\{t_i : t_1,\ ...,\ t_T\}$ and spatial locations $\{s_j : s_1,\ ...,\ s_n\}$, the observed

data is denoted $\mathbf{Y} = [\mathbf{y_1} \ ... \ \mathbf{y_n}]$, where $\mathbf{y_j}$ is a $T \times 1$ column vector containing data for location s_j. We will continue to use i to denote a temporal index and j to denote a spatial index throughout this work. $\mathbf{Y}$ may correspond either to the number of new infections or the number of individuals removed from the infectious population at each location/time pair, unknown quantities respectively denoted by the identically indexed $T \times n$ matrices: $\mathbf{I}^*$ and $\mathbf{R}^*$. A complete data model is specified by an appropriate distribution g with parameter vector Θ. For example, $\{y_{ij} \ | I^*_{ij}\} \overset{ind}{\sim} \ g(I^*_{ij}, \Theta)$ for $i = 1, ..., T; \quad j = 1, ..., n$. This structure can take innumerable forms, including identity, overdispersion, and binomial proportion. The choice of data model is ultimately dictated by the data used, constrained of course by the software available and parameter identifiability.

2.3 *Temporal Process Model*

The temporal structure employed by spatial SEIR models is their eponymous trait. $\mathbf{S}$,$\mathbf{E}$, $\mathbf{I}$, and $\mathbf{R}$ are all $T \times n$ matrices, and contain the unknown count parameters corresponding to the susceptible, exposed, infectious, and removed compartments, respectively. Their temporal relationship is described as a set of difference equations: $\mathbf{S}_{i+1} = \mathbf{S}_i - \mathbf{E}^*_i$; $\mathbf{E}_{i+1} = \mathbf{E}_i - \mathbf{I}^*_i + \mathbf{E}^*_i$; $\mathbf{I}_{i+1} = \mathbf{I}_i - \mathbf{R}^*_i + \mathbf{I}^*_i$; $\mathbf{R}_{i+1} = \mathbf{R}_i + \mathbf{R}^*_i$, subject to the constraint that $\mathbf{S}_i + \mathbf{E}_i + \mathbf{I}_i + \mathbf{R}_i = \mathbf{N} \ \forall \ i$. Here, $\mathbf{N}$ is the vector of fixed population sizes and i again denotes a particular time unit t_i.

The elements of the transition matrices, denoted above with asterisks, capture the number of individuals transitioning into each compartment of the same label. E^*_{ij}, for example, defines the newly exposed individuals at time t_i and spatial location s_j. These components are assigned the following chain binomial structure: $\{E^*_{ij} | \pi^{(SE)}_{ij}, S_{ij}\} \overset{ind}{\sim} \ \text{binom}(S_{ij}, \pi^{(SE)}_{ij})$, $\{I^*_{ij} | \pi^{(EI)}, E_{ij}\} \overset{ind}{\sim} \ \text{binom}(E_{ij}, \pi^{(EI)})$, and $\{R^*_{ij} | \pi^{(IR)}, I_{ij}\} \overset{ind}{\sim} \ \text{binom}(I_{ij}, \pi^{(IR)})$.

2.4 *Spatial Process Model*

The exposure probabilities, $\{\pi^{(SE)}_{ij}\}$, provide a clear way to incorporate spatial heterogeneity. They describe a combination of pathogen infectivity and population mixing. This structure

must therefore capture the relationship of the pathogen and population to any predictor variables as well as any spatial heterogeneity. Here, we briefly note the motivating assumptions and chosen spatial parameterization; additional details may be found in Web Appendix 1.1.

Consider the process by which people become infected with a communicable disease. Namely, consider the situation in which a person 'A' has contacted another person, 'B', who is infectious (for some suitable definition of contacted). Let p be the probability that person 'A' becomes infected with the disease, and let $q = 1 - p$. Now we introduce a number of assumptions. First, assume that the number of 'contacts' K_{ij} between a person of interest and other individuals within a spatial unit s_j at time t_i follows a Poisson distribution, $K_{ij} \sim \text{Pois}(\lambda_{ij})$. Second, when individuals travel to other spatial locations, their contact behavior is assumed to be well modeled by the contact behavior of that spatial unit. Finally, let contact between spatial locations be proportional to some known function $f(d_{jl})$, where d_{jl} is a specified distance metric between spatial locations s_j and s_l.

Define δ_{ij} to be the proportion of persons who are infectious at time t_i in spatial unit s_j, and p to be the probability an individual becomes infected given an epidemiologically significant contact. Then, letting $\text{Inf}(t_i, s_j)$ denote the event that a person in spatial unit s_j becomes infected at time t_i, we can derive the probability

$$P(\text{Inf}(t_i, s_j)) = \pi_{ij}^{(SE)} = 1 - \exp\left[-\delta_{ij} e^{\theta_{ij}} - \sum_{\{l \neq j\}} (f(d_{jl})\delta_{il} e^{\theta_{il}})\right]^{h_i} \quad (1)$$

where $\theta_{ij} = \log(\lambda_{ij} p)$ is the exposure intensity parameter for time t_i and location s_j, a reparameterization required for identifiability. The temporal offset, h_i, captures the relative length of continuous time over which the events are accumulated.

This approach immediately generalizes to the case in which contact between spatial locations depends on more than one distance parameter by relaxing the assumption that contact between spatial locations is proportional to a single distance metric. To strike a balance between flexibility and simplicity, each distance measure of interest is specified as an n by

n matrix, defining the set $\{\mathbf{D}_z : z = 1,\ ...,\ Z\}$ with corresponding spatial autocorrelation parameters $\{\rho_z\}$ subject to $\sum_{z=1}^{Z} \rho_z \leqslant 1$ and $\{0 \leqslant \rho_z < 1 : z = 1, ..., z\}$. This formulation gives rise to the exposure probability

$$\pi_{ij}^{(SE)} = 1 - \exp\left[\left\{-\eta_{i\cdot} - \sum_{z=1}^{Z} \rho_z(\mathbf{D}_z\eta_{i\cdot})\right\}_j^{h_i}\right],$$

where $\eta_{i\cdot} = \{\delta_{i1}e^{\theta_{i1}},\ ...,\ \delta_{in}e^{\theta_{in}}\}$. By appropriately defining 'distance' matrices, this spatial probability structure provides innumerable configurations, including, but not limited to, the usual Conditionally Auto-Regressive (CAR) model class of spatial dependence structures (Banerjee et al., 2004). Of course, one must be careful to ensure that the column spaces of the chosen matrices are distinct to maintain the identifiability of the autocorrelation terms.

2.5 *Parameter Model*

While most of the aforementioned parameters have natural prior distributions, there are a few structural notes worth discussing. First and foremost, the intensity process described by the set of parameters $\boldsymbol{\theta}$, where $[\boldsymbol{\theta}]_{ij} = \{\theta_{ij}\}$, is often of great interest to researchers. Estimation of a distinct parameter for each spatial location and time point is impossible, but a linear predictor prior structure, $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$, provides an intuitive and flexible lower dimensional representation for the intensity process. This structure can incorporate both time-varying and invariant covariates, as appropriate. An important special case is the single location SEIR model with a single intensity parameter. In our notation, such a model is constructed using just an intercept term in the linear predictor. Additional examples of this structure are given in the case studies in Section 5.

Specifying the parameter models for $\pi^{(EI)}$ and $\pi^{(IR)}$, introduced in the temporal process model, is relatively straightforward given that the time spent in the latent and infectious period may be reasonably modeled as a property of the pathogen rather than the population for many diseases. As in Lekone and Finkenstädt (2006), the set of E to I and I to R transition probabilities, $\pi_i^{(EI)}$ and $\pi_i^{(IR)}$ are given by $1-\exp(-h_i\gamma_{(EI)})$ and $1-\exp(-h_i\gamma_{(IR)})$,

respectively. In these expressions, $\gamma_{(EI)}$ is the rate at which individuals transition to the infectious from the exposed category, $\gamma_{(IR)}$ is the rate at which infectious individuals are removed, and, as before, h_i is a temporal offset. In the interest of clarity, for the remainder of this work we consider only the special case in which all of the $\{h_i\}$ are equal to one, although the work is trivially adapted to the inhomogeneous time case. This parameterization corresponds to mean latent and infectious periods $1/\gamma_{(EI)}$ and $1/\gamma_{(IR)}$ respectively, with exponentially distributed compartment membership time.

To ensure both flexibility and a proper range for the probability terms, independent gamma priors are placed on the respective γ terms, parameterized by $(\alpha_{(EI)}, \beta_{(EI)}, \alpha_{(IR)},$ and $\beta_{(IR)})$. As the average latent and infectious periods are well studied for many pathogens, we frequently have quite good information about their average values and range. The choice of appropriately informative priors is therefore advised, and can be easily accomplished by comparing the resulting compartment membership time to established ranges for a pathogen of interest.

2.6 *Implementation*

These models may be implemented using standard Markov chain Monte Carlo techniques, and in particular are well suited to Metropolis-Hastings and slice sampling. On the other hand, the large latent space and rich parameterization often result in high autocorrelation among samples. During the development of this work, which involved the creation of the open source libSpatialSEIR software package, we had success in autocorrelation reduction using a combination of tailored proposal distributions, decorrelation steps (Graves et al., 2011), and alternating joint sampling techniques (Brown, 2014).

3. The Basic Reproductive Number and Associated Quantities

3.1 *Introduction*

Consider the simplest special case of the preceding compartmental framework: a single location stochastic SEIR model with a single intensity parameter. The $T \times 1$ intensity process covariate matrix for this model, $\mathbf{X}$, consists of all ones, and has an associated scalar regression parameter β. The basic reproductive number for this model has been shown to be equal to the simple expression: $e^{\beta}/\gamma_{(IR)}$ (Lekone and Finkenstädt, 2006; Jones, 2007). This quantity makes intuitive sense, for in the notation used here, e^{β} captures the infection rate, which combines the contact behavior and infectivity of the pathogen, while $1/\gamma_{(IR)}$ gives the average number of time units during which a person remains infectious.

Nevertheless, there are problems with this common approach. When generalized to more richly parameterized intensity processes, this derivation requires either the choice of a meaningful 'baseline', or the calculation of many different context specific reproductive numbers. The intervention model of Lekone and Finkenstädt (2006), for example, estimates a baseline intensity parameter followed by a linear time component beginning on the date a significant intervention was launched. Such an approach (called by the authors $\mathcal{R}_0(t)$) has the benefit of not necessarily requiring a baseline or reducing complex behavior to a single number, but can still have odd consequences in real-world scenarios. For example, if a government quarantine goes into effect at time t_a, and is modeled by the indicator $1_{t_i > t_a}, i = 0, ..., T$, an individual infected at time t_{a-1} may have a drastically different estimated $\mathcal{R}_0(t)$ than a patient infected at time t_{a+1}, even though the vast majority of the first patient's infectious period will be spent under the quarantine. Conversely, should epidemic intensity worsen unexpectedly, reproductive number estimates at time periods immediately preceding the intensification will be artificially low. Note that $\mathcal{R}_0$ is a single parameter special case of $\mathcal{R}_0(t)$. We use both terms, depending on whether the emphasis is temporal or conceptual.

Another generalization used by the authors is the 'effective reproductive number', which is simply $\mathcal{R}_0(t)$ scaled by the proportion of susceptibles in a population. Such a measure might provide improved estimates for shrinking susceptible populations, but does not include the estimated infection size, and so only partially adapts to the population under study.

In the event that important covariates are unknown to the modeler (for example: unobserved changes in disease awareness in the population) none of these reproductive number estimates are capable of overcoming underspecification of the intensity process. To a greater or lesser degree, all are dependent on its parametric form. As we will demonstrate, this can result in heavily biased and obviously unreasonable estimates. Thus, while such definitions certainly serve a useful purpose, they are necessarily removed from the underlying estimated disease dynamics. We propose a more flexible estimate of the reproductive rate.

3.2 *The Empirically Adjusted Reproductive Number*

To define a measure of the reproductive rate in an actual population, we need only consider the expected number of secondary infections produced by a single infected individual in that population. Fortunately, this quantity has a natural and intuitive representation in the stochastic spatial SEIR framework.

Note that the average number of infections caused by a single individual from a given location at a particular time point is simply the total number of such infections divided by the number of infectious individuals. In the event that there are no infectious individuals, the average number of secondary infections is defined to be zero. Let the indicator $I_k(t_i, s_j, s_l)$ denote the event that a person k from spatial location s_j is infected at time t_i by contact from within spatial location s_l, and note that $P(I_k(t_i, s_j, s_l)) = 0$ unless person k is a member of the susceptible class $\mathcal{S}$. The expected number of such infections is then

$$E\big[\sum_{k=0}^{N_{ij}}(I_k(t_i, s_j, s_l))\big] = S_{ij} \cdot P(I_k(t_i, s_j, s_l)|k \in \mathcal{S})$$

and the average per infectious individual at (t_i, s_j) is just $\frac{S_{ij} \cdot P(I_k(t_i, s_j, s_l)|k \in \mathcal{S})}{I_{ij}}$. In the single

distance metric case, the associated contribution to the overall probability $\pi_{ij}^{(SE)}$ has a very simple simple form: $P(I_k(t_i, s_j, s_l)|k \in \mathcal{S}) = 1 - \exp\left\{-f(d_{jl})\delta_{il}e^{\theta_{il}}\right\}$. The general case, for which we must consider the contribution from each distance metric, can be written as $1 - \exp\left(-\sum_{z=1}^{Z} \rho_z \left\{\mathbf{D}_z\right\}_{jl} \cdot \eta_{il}\right)$, where $\eta_{il} = \delta_{il}e^{\theta_{il}}$. As before, the contact events are considered independent and the $\{\mathbf{D}_z\}$ define the set of distance matrices.

These components, which capture the spatial structure of expected secondary infections, can be arranged into a matrix which is a single time unit analogue of the $n \times n$ 'next generation matrix' $\mathbf{G}(t_i)$ (Allen and van den Driessche, 2008). Expressions for the elements, $G_{jl}(t_i)$, of these matrices are given by pre-multiplying the expressions for the single and multiple metric cases by the ratio, $\frac{S_{ij}}{I_{il}}$.

The usual next step in this approach to basic reproductive number estimation is to calculate the dominant eigenvalue, or spectral radius, of the next generation matrix (Allen and van den Driessche, 2008). However, more information can often be gleaned from careful examination of the row sums of the constructed matrix, which give the average number of infections caused by each infectious individual in the spatial location with the same index at that time point. To generalize this result to the lifetime of the pathogen, we compute the expected total number of such infections over time: $\sum_{t=t_i}^{t_\infty} \mathbf{G}(t) \cdot \left[\prod_{k=t_{j+1}}^{t}(1 - \pi_k^{(IR)})\right]$.

While $\mathcal{R}^{(EA)}$ applies to the study population rather than a hypothetical susceptible population, a similar thresholding argument to that used for $\mathcal{R}_0$ applies. In the language of Heffernan et al. (2005), if $\mathcal{R}^{(EA)}$ is greater than one, the pathogen is expected to further colonize the population under study, while the opposite is true for $\mathcal{R}^{(EA)}$ less than one.

3.3 $\mathcal{R}_0$ *as a Special Case*

Consider the derived expression for $\mathcal{R}^{(EA)}(t)$ applied to an approximation of the hypothetical population evoked by $\mathcal{R}_0$. That is, consider a single spatial unit with a fixed size population of susceptibles at the time t_i when a single infectious individual is introduced. Note that, in this

case, $S_t \approx N \; \forall \; t$. If we restrict our attention to the simple single parameter intensity function and assume equally spaced temporal intervals, an interesting result may be obtained by letting the population size become arbitrarily large ($N \to \infty$). In this setting, the expression for $\mathcal{R}^{(EA)}$can be shown to converge to $e^{\theta}/\pi^{(IR)}$ (Web Appendix 1.2).

Therefore, by imposing the aforementioned constraints on the infection events and reintroducing the hypothetical population employed by $\mathcal{R}_0$, we find that $\mathcal{R}^{(EA)}$ is analogous to the traditional $\mathcal{R}_0$ estimate with $1/\gamma_{(IR)}$, the average infectious period based on an exponential time assumption, replaced by $1/\pi^{(IR)}$, the average infectious period implied by the constant transition probability (geometric). This correspondence provides an insight into some of the information which is lost in the traditional approach. In particular, $\mathcal{R}_0(t)$ ignores the nonlinear effect on the contact rate of the number of infectious individuals.

3.4 *Estimation*

The formulation of the empirically adjusted reproductive number allows sampling via MCMC algorithms, but invites the question of how to perform the requisite infinite summation in practice. Fortunately, in realistic epidemics, the pathogen lifespan weighting term, $\prod_{k=t_{j+1}}^{t}(1-\pi_k^{(IR)})$, will quickly and monotonically approach zero. In other words, patients are assumed to cease being infectious in finite time with probability one. Therefore, we may perform the summation forward in time until the probability of a person remaining infectious drops below a reasonable threshold. At the end of the study period, one can simply re-use the final available term, or employ estimates based on predicted values.

4. Methods

4.1 *Case Study: 1995 Ebola Outbreak in Kikwit*

Kikwit is a large city in the Bandundu region of the Congo which was the epicenter of an Ebola outbreak in 1995 (Chowell et al., 2004). There were a total of 316 documented cases

of Ebola Virus Disease (EVD) between March and July of that year, and the epidemic has been well studied in the time since.

This data set encompasses a single location, and the canonical analysis employs a simple intensity process which incorporates an intercept up to the date on which intervention efforts began, and adds a linear temporal term after that date (Lekone and Finkenstädt, 2006). In order to examine the behavior of $\mathcal{R}_0$, effective $\mathcal{R}_0$, and $\mathcal{R}^{(EA)}$, we first perform this standard analysis and then examine an underspecified version incorporating only an intercept.

The original study fixed the E to I and I to R transition parameters at 1/5 and 1/7 respectively, while we set their prior mean equal to these values with high precision (1000 equivalent samples). In addition, the authors model the intensity intercept on the log scale relative to our linear predictor approach. We therefore employ a Gaussian, rather than gamma prior structure to ensure it has a properly signed contribution to the intensity process.

Three MCMC samplers were started from random parameter values and samples were drawn until convergence. This was established by requiring that the Gelman and Rubin diagnostic (Gelman and Rubin, 1992) was less than or equal to 1.02 for all model parameters.

4.2 *Case Study: 2014 Ebola Outbreak in West Africa*

The 2014 and 2015 West African Ebola epidemic has captured international attention for its unprecedented size and duration as well as the sheer scope of the human tragedy. While a small number of cases have been transported to or contracted in Nigeria, Mali, the United States, Senegal and Spain, the core of epidemic activity has been the three neighboring nations of Guinea, Liberia, and Sierra Leone. The first recorded case was observed in Guinea in December 2013, though the epidemic began to more rapidly spread in March and April of 2014 (Frieden et al., 2014). As such, our analyses were restricted to the three countries most affected by the epidemic: Guinea, Liberia, and Sierra Leone.

Unlike the single location analysis, this epidemic is still underway at the time of manuscript

preparation and thus does not have the benefit of hindsight in evaluating interventions or knowledge of the time disease spread will conclude. Moreover, local public health infrastructure has proven inadequate to contain, let alone defeat, the epidemic (Fauchi, 2014). This environment has produced incredible challenges for national and international organizations tasked with fighting the disease, and as a result has complicated modeling efforts due to the difficulty of obtaining reliable data (Farrar and Piot, 2014). Despite these challenges, the situation provides a rigorous test of analysis techniques. In particular, we focus on reproductive number estimation in the face of uncertainty both in terms of actual infection counts and the proper parametric form of the intensity process; while information exists on intervention efforts, we have found no complete and thorough accounting of all such active programs with sufficient spatial and temporal resolution to construct such a model.

Incidence data is available from the World Health Organization's situation report publications, and is modeled under overdispersion to partially account for the uncertainty involved. The original data aggregation was performed by a community of volunteers using the GitHub social coding site (Rivers et al., 2014); online collaboration and rapid data analytics have been persistent features of this epidemic. Due to the uneven temporal availability of incidence reports between countries, these data required further processing prior to use in the spatial SEIR framework described above. Records were aggregated to to ensure that a minimum of one week passed between incidence reports. This had the additional effect of smoothing some of the observed short term variability in new cases. Given the limited and changing availability of treatment beds, the potentially shifting geographic reach of surveillance teams, and the internationally coordinated data collection efforts, this short term variability could plausibly reflect surveillance patterns more than actual changes in transmission.

We begin this set of analyses by invoking a simple intercept model in which each of the three nations is assumed to have a separate, constant, intensity value. This simple structure

is compared to several expanded versions which include temporal basis splines of varying degrees of freedom, produced with the `splines` R package (R Core Team, 2013). This approach allows us to construct flexible models in the absence of structural information about the Ebola epidemic which would otherwise help inform the intensity process, such as finer spatiotemporal indexing, data on funerals, and comprehensive data on public health efforts, among other things.

Each set of nations was given a separate spatial autocorrelation parameter to examine potential differences in cross border spread. In our experience, especially in the simple intercept case, such a flexible spatial process may cause parameter identifiability issues if the elements of $\boldsymbol{\rho}$ are given flat priors and the $\boldsymbol{\beta}$ parameters are given mean zero normal priors; a pair of nations with similar average intensity parameters may trade spatial correlation for local intensity. This behavior is easily addressed by placing a beta prior on the spatial autocorrelation parameters, $\boldsymbol{\rho}$, which restricts them to a reasonable range.

4.3 *Simulation*

Examination of the behavior of the empirically adjusted reproductive number under simulation is not so straightforward as it may initially appear. The underlying quantity we wish to estimate is the true number of secondary infections in a particular context, however data simulated from the population averaged models described above only permits the calculation of the expected value of such counts. Our simulation work thus falls into two categories: agent based simulations comparing the ability of $\mathcal{R}^{(EA)}$, $\mathcal{R}_0(t)$, and effective $\mathcal{R}_0(t)$ to estimate the true number of context specific secondary infections under correct and misspecified intensity processes, and population averaged simulations concerning the estimation of the expected counts. We also consider $\mathcal{R}^{(EA)}$ as an approximation of $\mathcal{R}_0(t)$. We find that our method provides improved estimation of the true number of secondary infections, and can

be profitably compared to traditional reproductive number estimates. These simulations are described in detail in Web Appendix 2, where results are also presented.

5. Results

5.1 *Case Study: 1995 Ebola Outbreak in Kikwit*

Figure 1 illustrates basic and empirically adjusted reproductive number curves for both reasonable and underspecified intensity processes. In the rightmost graphs, corresponding to the intervention model, note that both measures follow a similar trajectory. No comparable graphic is available in Lekone and Finkenstädt (2006), although the associated $\mathcal{R}_0$ estimates are similar and well within the given credible intervals. Effective $\mathcal{R}_0$ is not shown, being entirely indistinguishable from $\mathcal{R}_0(t)$ as $\frac{S_i}{N} \approx 1 \; \forall \; i$.

More interestingly, the leftmost graphs demonstrate both the difference between these measures in underspecified models and the sensitivity of $\mathcal{R}_0$ estimation to the intensity process. The measure of $\mathcal{R}_0(t)$ is, naturally, completely linear as it is a reflection of the parametric form of the intensity process (i.e., an intercept). Moreover, we see that it is clearly biased downward; the literature is in agreement that $\mathcal{R}_0$ for this epidemic was between 1.36 and 1.83 (Lekone and Finkenstädt, 2006).

$\mathcal{R}^{(EA)}$ here distinguishes itself by shrinking towards zero after the intervention date, even though no intervention information was incorporated into the model. Of course, this method does not completely recover the reproductive trend given by the standard parameterization, but as discrepancy between the oversimplified model and the underlying disease dynamics widens, the measure is increasingly shrunk towards the true value. We again observe that the entire curve is still negatively biased, though to a lesser degree than the traditional measure.

Most interesting, perhaps, is the difference between the two measures. Such deviation indicates, as observed under simulation, that the observed epidemic behavior and the form

of the intensity differ, and should indicate to a modeler that important factors in the disease intensity process are not accounted for. This is especially useful given that both measures can be estimated as part of the usual model fitting process.

5.2 *Case Study: 2014 Ebola Outbreak in West Africa*

Examining the new case counts in Figure 2, we see that the temporal trend is not so obvious as in the earlier epidemic. In Web Appendix 2, comparisons of $\mathcal{R}^{(EA)}(t)$ and $\mathcal{R}_0(t)$ curves are presented for models with zero (intercept only), three, and five degree of freedom cubic basis splines incorporated into their intensity processes. Based on our previous observation that well specified models tend to exhibit reproductive curves of similar shape, we conclude that the three degree of freedom spline basis strikes the best balance between model fit and parsimony. Clearly, additional work remains to quantify and characterize such a selection procedure, however the comparison remains valuable in the face of uncertainty.

Posterior quantiles for the selected final model are available in Web Appendix 2, though a few key features are noted here. Although there is some overlap in the posterior credible regions for the three country specific intercepts, Sierra Leone had the highest median epidemic potential, followed closely by Liberia. Guinea had a substantially lower estimated intensity intercept. Interpretation of the three temporal basis parameters is not so straightforward, though only the last had a 95% credible region which did not include zero. Illustration of the values of these components over time is also available in the Web Appendix, both for estimation and prediction. Interpretation of credible regions for predicted quantities is complicated by the arbitrary nature of basis splines; we have not quantified our considerable uncertainty about their functional form.

The three included spatial parameters capture contact between Guinea and Liberia, Guinea and Sierra Leone, and Liberia and Sierra Leone, respectively. All three parameters are distinctly nonzero, however the border between Liberia and Sierra Leone had by far the

highest estimated contact rate: two times higher than the other two borders. Given this observed heterogeneity, future research might explore an asymmetrical contact specification to examine the effect of any net population flows between countries.

Beyond parameter estimation, epidemic prediction is of great interest. To this end, one may run the simulation forward in time for as long as desired for each converged MCMC sample. Of course, longer term predictions are subject to additional uncertainty, which may not be fully accounted for in the absence of scientifically motivated intensity functions. Presented in Figure 3 are the predictions for the final selected model. Confidence bands have been omitted for clarity, and because they do not reflect our uncertainty about the form of the intensity process. Interpretation of the results must be cautious. In this case, we consider the result to provide weak evidence, as of January 2015, that the epidemic would continue to slow until it is containment in April or May of 2015. Subsequent analysis in April 2015 largely agreed with this trajectory, although different dynamics appear to have taken over as cases diminished (Web Appendix 2). While we lack the data to investigate such a possibility formally, this shift could plausably be attributed to unmodeled heterogeneity at a fine spatial scale.

6. Conclusions

Both $\mathcal{R}^{(EA)}$ and $\mathcal{R}_0(t)$ are intuitive quantities which capture the infectious behavior of pathogens. We have demonstrated that the conceptual differences, which include the applicable population and the approach to defining the average number of secondary infections, produce quite distinct behavior in practice. Moreover, we have demonstrated that this distinction can be used to motivate, albeit informally, model selection efforts. Based on this work, it is our opinion that traditional reproductive measures in stochastic compartmental models should not be applied and interpreted without first comparing their behavior with

this, more flexible, measure. This is particularly important given the observed tendency for underspecified intensity processes to result in unreasonable reproductive number estimates.

One of the primary applications of reproductive number estimation is the ability to at least partially address the question of whether a particular disease may invade an, as yet, unaffected population. With this in mind, some might be concerned that our empirically adjusted approach loses some of this generalizability. While it is certainly true that $\mathcal{R}^{(EA)}$ does not immediately provide a scalar summary quantity like $\mathcal{R}_0$, it is worth remembering that the basis for $\mathcal{R}_0$ estimation is the intensity process - a construct which addresses the confounded behavior of pathogen infectivity and population mixing. Generalizing such a measure requires careful expert input in any case, as population mixing behaviors may differ dramatically from region to region. Moreover, we have seen that incautious use of such a simplified measure may give very misleading results.

More work remains to characterize summary methods which may be applied to $\mathcal{R}^{(EA)}$, as well as to formalize the model selection process motivated by the parameter. Nevertheless, the utility of this method is clear, both as a realistic estimate of epidemic reproductive behavior and as a counterpoint to traditional methods.

7. Supplementary Materials

Web Appendices 1.1 ,1.2, and 2, referenced in Sections 2, 3, and 5 respectively, in addition to a companion R package and collection of scripts implementing all included analyses and simulations, are available with this paper at the *Biometrics* website on Wiley Online Library. All referenced data is also included.

Acknowledgement

This research was supported in part through computational resources provided by The University of Iowa, Iowa City, Iowa.

References

Allen, L. J. and van den Driessche, P. (2008). The basic reproduction number in some discrete-time epidemic models. *Journal of Difference Equations and Applications* **14,** 1127–1147.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Heirarchical Modeling and Analysis for Spatial Data.* Chapman & Hall.

Brown, G. D. (2014). *spatialSEIR: A spatial SEIR epidemic modeling framework.* R package.

Cauchemez, S., Carrat, F., Viboud, C., and Boëlle, P. (2004). A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statistics in Medicine* **23,** 3469–3487.

Cauchemez, S., Donnelly, C. A., Reed, C., Ghani, A. C., Fraser, C., Kent, C. K., et al. (2009). Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States. *New England Journal of Medicine.* **361,** 2619–2627.

Chis Ster, I. and Ferguson, N. M. (2007). Transmission parameters of the 2001 foot and mouth epidemic in Great Britain. *PLoS One* **2,** e502.

Chis Ster, I., Singh, B. K., and Ferguson, N. M. (2009). Epidemiological inference for partially observed epidemics: The example of the 2001 foot and mouth epidemic in Great Britain. *Epidemics* **1,** 21–34.

Chowell, G., Hengartner, N., Castillo-Chavez, C., Fenimore, P., and Hyman, J. (2004). The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *Journal of Theoretical Biology* **229,** 119–126.

Cook, A. R., Otten, W., Marion, G., Gibson, G. J., and Gilligan, C. A. (2007). Estimation of multiple transmission rates for epidemics in heterogeneous populations. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 20392–20397.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-temporal Data.* John Wiley & Sons,

Inc.

Deardon, R., Brooks, S. P., Grenfell, B. T., Keeling, M. J., Tildesley, M. J., Savill, N. J., et al. (2010). Inference for individual-level models of infectious diseases in large populations. *Statistica Sinica* **20,** 239–261.

Farrar, J. J. and Piot, P. (2014). The Ebola emergency - immediate action, ongoing strategy. *New England Journal of Medicine* **371,** 1545–1546.

Fauchi, A. S. (2014). Ebola - underscoring the global disparities in health care resources. *New England Journal of Medicine* **371,** 1084–1086.

Frieden, T. R., Damon, I., Bell, B. P., Kenyon, T., and Nichol, S. (2014). Ebola 2014 - new challenges, new global response and responsibility. *New England Journal of Medicine* **371,** 1177–1180.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7,** 457–472.

Graves, T. L., Speckman, P. L., and Sun, D. (2011). Improved mixing in MCMC algorithms for linear models. *Journal of Computational and Graphical Statistics* .

Heffernan, J., Smith, R., and Wahl, L. (2005). Perspectives on the basic reproductive ratio. *Journal of the Royal Society. Interface* **2,** 281–293.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review* **42,** 599–653.

Hooten, M. B., Anderson, J., and Waller, L. A. (2011). Assessing North American influenza dynamics with a statistical SIRS model. *Spatial and Spatiotemporal Epidemiology* **1,** 177–185.

Jandarov, R., Haran, M., Bjørnstad, O., and Grenfell, B. (2014). Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease. *Journal of the Royal Statistical Society* **63,** 423–444.

Jewell, C., Keeling, M., and Roberts, G. (2009). Predicting undetected infections during the

2007 foot-and-mouth disease outbreak. *Journal of the Royal Statistical Society Interface* **6,** 1145–1151.

Jones, J. H. (2007). Notes on $\mathcal{R}_0$. *Stanford University* .

Keeling, M. (2004). The implications of network structure for epidemic dynamics. *Theoretical Population Biology* **67,** 1–8.

Kermack, W. and McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society, London* **115,** 700–721.

LaBute, M. X., McMahon, B. H., Brown, M., Manore, C., and Fair, J. M. (2014). A flexible spatial framework for modeling spread of pathogens in animals with biosurveillance and disease control applications. *ISPRS International Journal of Geo-Information* **3,**.

Lekone, P. E. and Finkenstädt, B. F. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics* **62,** 1170–1177.

Porter, A. T. and Oleson, J. J. (2014). A multivariate CAR model for mismatched lattices. *Spatial and Spatio-temporal Epidemiology* **11,** 79–88.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rivers, C. et al. (2014). *Data for the 2014 Ebola Outbreak in West Africa*.

Sattenspiel, L. and Dietz, K. (1995). A structured epidemic model incorporating geographic mobility among regions. *Mathematical Biosciences* **128,** 71–91.

van Boven, M., Donker, T., van der Lubben, M., van Gageldonk-Lafber, R. B., te Beest, D. E., Koopmans, M., et al. (2010). Transmission of novel influenza A (H1N1) in households with post-exposure antiviral prophylaxis. *PLoS One* **5,** e11442.

Verdasca, J., da Gama, T., Nunes, A., Bernardino, N., Pacheco, J., and Gomes, M. (2005). Recurrent epidemics in small world networks. *Journal of Theoretical Biology* **233,** 553–561.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

Figure 1. 1995 Ebola Outbreak in Kikwit: Cases and Model Based Reproductive Numbers

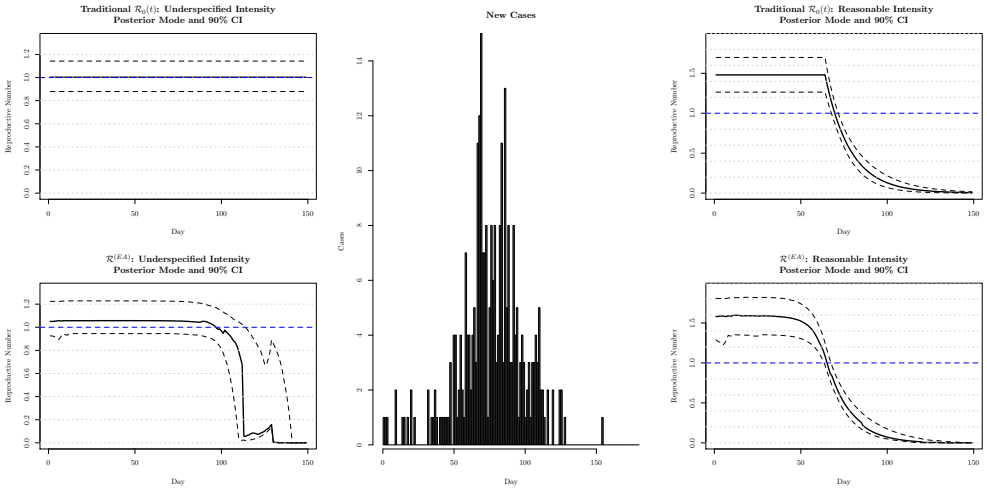

Figure 2. 2014 Ebola Outbreak in West Africa: Estimated Infections Per Day

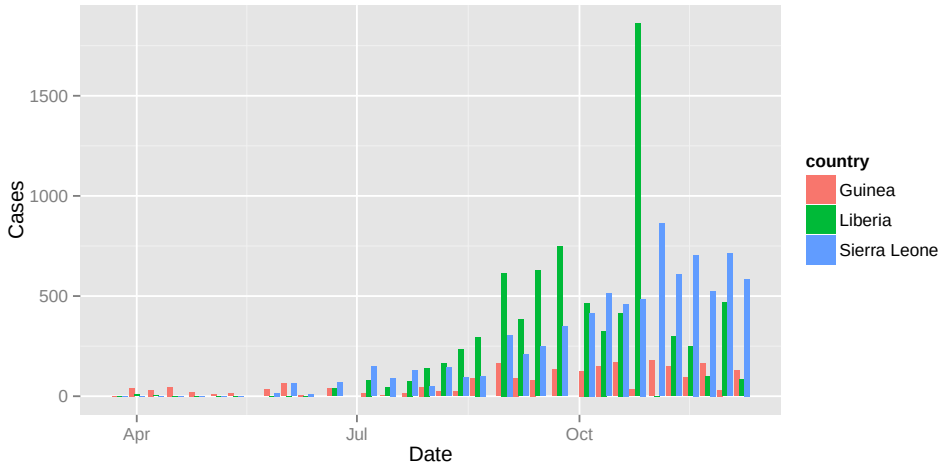

Figure 3. Long Term Infectious Count Predictions for the 2014 Outbreak: Three Degree of Freedom Model