Biometrics 000, 1–26 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 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: Compartmental model; Model selection; Underspecification; Ebola; SEIR; Spatial epidemiology.

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 the 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,* 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 chaotic, and are not always well summarized by a single number. Here we examine a different approach to estimating the reproductive characteristics of epidemics which may better 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, which includes the important single location special case.

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. As will be made apparent, however, our development is quite distinct, and the two measures do not share important mathematical properties.

In the following work, we first introduce the stochastic spatial SEIR model class; 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 introduce the derivation of $\mathcal{R}^{(EA)}$. Next, the conceptual and mathematical relationships between this new quantity and traditionally defined $\mathcal{R}_0$ are discussed. Finally, the practical implications of reproductive number choice are explored via two examples: the 1995 Ebola outbreak in the Congolese city of Kikwit, and the 2014 Ebola epidemic in West Africa. The first of these 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 the SEIR framework have been developed in the intervening years, though it is only recently that stochastic formulations have received a thorough spatial treatment. This is likely in part due to the computational difficulties inherent in fitting spatial SEIR models. One approach to reducing this challenge 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.

Much additional generalization has been done in the context of household models, where houses define population clusters and SIR or SEIR compartment structures provide a model for disease transmission (see Cauchemez et al. (2004, 2009); van Boven et al. (2010)). Such techniques generally consider epidemic spread within population clusters and with an

external community, whereas our approach focuses on pathogen spread within and between population clusters rather than with exogenous sources. Additional literature has examined the similar concept of implementing compartmental dynamics over social or contact networks. Verdasca et al. (2005) and Keeling (2004) explore this approach primarily from a simulation perspective, with a particular emphasis on relaxing the assumption of homogeneous mixing between individuals. Sattenspiel and Dietz (1995) instead 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. The stochastic work of Porter and Oleson (2014) more closely matches our spatial development.

Beginning in section 2.2, we describe the structure of the general stochastic spatial SEIR model class, following many standard conventions for the specification of hierarchical models (Cressie and Wikle, 2011). We introduce the compartmental structure, develop a data model to relate the observed information to the phenomenon of interest, define process models to describe the underlying spatial and temporal dynamics, and briefly discuss the parameter model which completes the specification.

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. These observations may correspond either to a measure of the number of new infections or the number of individuals removed from the infectious population

at each location/time pair. The associated model parameters for these two potential data types are the identically indexed $T \times n$ matrices: $\mathbf{I}^*$ and $\mathbf{R}^*$. In the first case, given an appropriate distribution g with parameter vector Θ and information on newly infectious members of a population, a data model would be specified by $\{y_{ij} \mid I^*_{ij}\} \overset{ind}{\sim} g(I^*_{ij}, \Theta)$ for $i = 1, ..., T; \quad j = 1, ..., n$.

This structure can take a number of forms. For example, if counts of newly infectious individuals are observed with error, an overdispersion model might be appropriate. Alternatively, it may be more reasonable to assume that a binomial proportion of cases are observed. 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 $\mathbf{S}_i + \mathbf{E}_i + \mathbf{I}_i + \mathbf{R}_i = N \,\forall\, i$. Here, 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 astrices, are assigned the following chain binomial structure:

$$\begin{aligned} \{E^*_{ij} | \pi^{(SE)}_{ij}, S_{ij}\} &\overset{ind}{\sim} \ binom(S_{ij}, \pi^{(SE)}_{ij}) \\ \{I^*_{ij} | \pi^{(EI)}, E_{ij}\} &\overset{ind}{\sim} \ binom(E_{ij}, \pi^{(EI)}) \\ \{R^*_{ij} | \pi^{(IR)}, I_{ij}\} &\overset{ind}{\sim} \ binom(I_{ij}, \pi^{(IR)}) \end{aligned} \tag{1}$$

The exposure probabilities, $\{\pi^{(SE)}_{ij}\}$, provide an ideal means to incorporate spatial heterogeneity, a generalization addressed in section 2.4. Both the $\{\pi^{(EI)}\}$ and $\{\pi^{(IR)}\}$ parameters are principally functions of the pathogen of interest, and are discussed in section 2.5.

2.4 *Spatial Process Model*

The parameters associated with the set $\left\{\pi_{ij}^{(SE)}\right\}$ describe the intensity process: 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:

- The number of 'contacts' K between a person of interest and other individuals within a spatial unit s_j at time t_i follows a Poisson distribution, $K \sim Pois(\lambda_{ij})$.
- When individuals travel to other spatial locations, their contact behavior is well modeled by the contact behavior of that spatial unit.
- Contact between spatial locations is 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 $Inf(t_i, s_j)$ denote the event that a person becomes infected from contact within spatial unit s_j at time t_i, we can derive the probability

$$P(Inf(t_i, s_j)) = \pi_{ij}^{(SE)} = 1 - exp\left\{-\delta_{ij}e^{\theta_{ij}} - \sum_{\{l \neq j\}} (f(d_{jl})\delta_{il}e^{\theta_{il}})\right\} \quad (2)$$

where $\theta_{ij} = log(\lambda_{ij}p)$, the exposure intensity parameter for t_i, s_j. Note that this common reparameterization is required due to the unidentifiability of λ_{ij} and p.

This approach immediately generalizes to the case in which contact between spatial loca-

tions 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

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

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 the usual Conditionally Auto-Regressive (CAR) model class of spatial dependence structures. For example, in the case where spatial data are indexed by discrete areal units, a single neighborhood matrix $\mathbf{D}$ with $D_{jl} = 1$ if spatial units s_j and s_l are neighbors and 0 on the diagonal and for non-neighbors gives rise to a CAR model (Banerjee et al., 2004). On the other hand, one could define a more traditional distance metric between discrete point locations (ie, city centroids) to allow continuous spatial dependence. In the extreme, $\binom{n}{2}$ sparse matrices could be used to specify a separate spatial correlation parameter for each pair of locations. Indeed, because the formulation does not require a single metric, several of these approaches may be used in tandem. Of course, one must always be careful to ensure that the column spaces of the chosen matrices are distinct to maintain the identifiability of the autocorrelation terms $\{\rho\}$.

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 $\{\theta_{ij}\}$ is often of great interest to researchers. As with the other spatiotemporally indexed parameters, we organize the set of $\{\theta_{ij}\}$ into a $T \times n$ matrix $\boldsymbol{\theta}$. 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.

The log link, shown in (2), ensures that the contribution of the intensity parameters is non-negative, while an appropriately structured covariate matrix $\mathbf{X}$ and parameter vector $\boldsymbol{\beta}$ provide a flexible intensity structure incorporating 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 parameter 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 (1), is relatively straightforward given that the time spent in the latent and infectious period may reasonably be considered a property of the pathogen rather than the population for most diseases. As in Lekone and Finkenstädt (2006), the set of E to I and I to R transition probabilities is given by

$$\pi_i^{(EI)} = 1 - exp(-h_i\gamma_{(EI)}), \quad \pi_i^{(IR)} = 1 - exp(-h_i\gamma_{(IR)}). \tag{3}$$

In (3), $\gamma_{(EI)}$ 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 h_i is a temporal offset which captures the relative length of continuous time over which the events of time t_i are accumulated. 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, gamma priors are placed on the respective γ terms. As these parameters are properties of the pathogens, rather than populations, we generally have quite good information about their average values and range. The choice of informative hyperparameters $(\alpha_{(EI)}, \beta_{(EI)}, \alpha_{(IR)}, \beta_{(IR)})$ is 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. More implementation details, tutorials, examples and source code are available online (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. In the notation established in section 2, such a model defines the $T \times 1$ compartment matrices $\mathbf{S}, \mathbf{E}, \mathbf{I}, \mathbf{R}$, and corresponding transition matrices $\mathbf{E}^*, \mathbf{I}^*$ and $\mathbf{R}^*$. Additionally, the $T \times 1$ intensity process covariate matrix $\mathbf{X}$ consists of all ones, and has an associated scalar regression parameter β. In this case, therefore, $\boldsymbol{\theta}$ is entirely determined by this single parameter.

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 clearly requires either the choice of a meaningful 'baseline', or the calculation of many different context specific reproductive numbers. An illustrative intensity process generalization is presented by the intervention model of Lekone and Finkenstädt (2006), which 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 $R_0(t)$) has the benefit of not necessarily requiring a baseline or reducing complex behavior to a single number, but can still have odd consequences when interpreted in real-world scenarios. For example, if a well enforced 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 R_0 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 throughout this work, 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 estimated number of susceptibles in a population. Such a measure might be expected to give reasonable estimates for shrinking susceptible populations, but neither incorporates the estimated infection size nor gives an estimate of traditional $\mathcal{R}_0$ which, again, posits an entirely susceptible population.

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*

Consider again the informal definition of the basic reproductive number: the expected number of secondary infections produced by a single infected individual in an entirely susceptible population. A first step towards defining a measure of the reproductive rate in an actual population is to relax the clause requiring "an entirely susceptible population", and instead compute the expected number of secondary infections during the lifetime of the pathogen over the observed epidemic, a quantity with a natural and intuitive representation in the stochastic spatial SEIR framework. The informal definition might initially appear to evoke the 'effective reproductive number' of Chowell et al. (2004), or the replacement number of Hethcote (2000), however the resulting development will demonstrate that neither completely incorporates the exposure and contact behavior implied by the model structure.

Note first 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 infections caused under such conditions 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

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

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

distance metric case, the associated contribution to the overall probability $\pi_{ij}^{(SE)}$ has a 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

$$P(I_k(t_i, s_j, s_l)|k \in \mathcal{S}) = 1 - exp\left(-\sum_{z=1}^{Z}\rho_z\left\{\mathbf{D}_z\right\}_{jl}\cdot\eta_{il}\right),$$

where $\eta_{il} = \delta_{il}e^{\theta_{il}}$. As before, the contact events are independent and the $\{\mathbf{D}_z\}$ define the set of distance matrices.

These expectations 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 Dreissche, 2008), by averaging the infection counts over the number of responsible infectious individuals. 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: $\frac{S_{ij}}{I_{il}}$. Additionally, recall the diagonal case ($j = l$), where $d_{jj} = 0$ and $f(0) = 1$:

$$G_{ii}(t_i) = \frac{S_{ij}}{I_{ij}}\cdot\left[1 - exp\left\{-\delta_{ij}e^{\theta_{ij}}\right\}\right] = \frac{S_{ij}}{I_{ij}}\cdot\left[1 - exp\left\{-\delta_{ij}e^{\theta_{ij}}\right\}\right].$$

The usual next step in the next generation matrix approach to basic reproductive number estimation is to calculate the dominant eigenvalue, or spectral radius (Allen and van den Dreissche, 2008), however more information can often be gleaned from careful examination of the rows and associated row sums of the constructed matrix. As defined, each cell describes a different component of the estimated disease transmission dynamics for the time at which the matrix is calculated. The row sums of the matrix 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 take the expected value over the distribution of infection times; i.e. take the sum the resulting aggregated values over time, weighted by the probability that an individual joining the infectious class at time t is

still a member:

$$\sum_{t=t_i}^{t_\infty} \mathbf{G}(t) \cdot \left[\prod_{k=t_{j+1}}^{t} (1 - \pi_k^{(IR)}) \right].$$

This approach might, superficially, appear to be akin to applying a smoothing operation to the $\mathcal{R}_0(t)$ estimator defined by Lekone and Finkenstädt (2006), however it allows a much more flexible fit given its dependence on the actual estimated compartment sizes rather than the parametric form of the intensity process alone. Such differences will be emphasized in the results section.

While $\mathcal{R}^{(EA)}$ applies to the population under study rather than a hypothetical entirely susceptible population, the same thresholding argument used for $\mathcal{R}_0$ applies. In the language used by Heffernan et al. (2005), if $\mathcal{R}^{(EA)}$ is greater than one, the pathogen may be expected to further colonize the population under study, and when $\mathcal{R}^{(EA)}$ is less than one we expect the disease to be cleared. There are also instructive mathematical connections between $\mathcal{R}^{(EA)}$ and $\mathcal{R}_0$.

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 invoked 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, $S_t \approx N \,\forall\, t$. For clarity, consider again the simple single parameter intensity function and assume equally spaced temporal intervals. This gives the corresponding expression for the proposed reproductive number estimate,

$$\mathcal{R}^{(EA)}(t_i) = \sum_{t=t_i}^{t_\infty} \left(\frac{N}{I_t} \right) (1 - exp\{-\frac{I_t}{N} e^{\theta}\})(1 - \pi^{(IR)})^{(t-t_i)}.$$

If we make the further assumption that I_t remains equal to one long enough that the remaining terms in this infinite summation are negligible, we have the approximate equality

$$\mathcal{R}^{(EA)}(t_i) \approx \left[\frac{(1 - exp\{-\frac{1}{N} e^{\theta}\})}{\left(\frac{1}{N}\right)} \right] \left[\sum_{t=t_i}^{t_\infty} (1 - \pi^{(IR)})^{(t-t_i)} \right]. \tag{4}$$

This assumption is often not reasonable, and is a primary driver of the difference between the behavior of $\mathcal{R}^{(EA)}$ and $\mathcal{R}_0$. It can be written slightly more explicitly by letting $t_f,\ f > i$ be the temporal index at which the first exposed individual becomes infections, and specifying that

$$0 \approx \sum_{t=t_f}^{t_\infty} \left(\frac{N}{I_t}\right) (1 - exp\{-\frac{I_t}{N} e^{\theta}\})(1 - \pi^{(IR)})^{(t-t_i)},$$

which is only reasonable if $t_f \gg t_i$ or $1 - \pi^{(IR)}$ is very small.

As the population size becomes arbitrarily large ($N \to \infty$), the expression in (4) 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 uncover a compelling connection. In this case, the $\mathcal{R}^{(EA)}$ method 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. Based on this difference, we might reasonably expect $\mathcal{R}^{(EA)}$ to be higher than $\mathcal{R}_0(t)$ for active epidemics.

3.4 *Estimation*

All of the reproductive number estimates described thus far provide point estimates of their respective threshold quantities. Clearly, both an idea of the reliability of such estimates as well as appropriate credible intervals are necessary for a complete analysis. Conveniently, all of the estimates are simple functions of the variables sampled under MCMC and can therefore be computed using standard Bayesian estimation techniques upon the establishment of chain convergence.

The formulation of the empirically adjusted reproductive number invites an additional

question, however: how to perform this infinite summation in practice. Fortunately, realistic epidemics (those to which the SEIR approach is at least somewhat applicable), 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, it is sufficient to perform the summation forward in time until the probability of a person remaining infectious drops below a reasonable threshold (ie, some multiple of machine epsilon). Should the sum approach the end of the study period before this occurs, one can make the assumption that the model state at the end of the study is a reasonable approximation of the period following, and simply re-use the final available term. In fact, the modeler may use the current parameter and compartment values to make a prediction of appropriate length and base the estimate on the predicted values. As before, this may be implemented using the output of converged MCMC chains.

4. Analysis 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; Lekone and Finkenstädt, 2006). 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. In particular, the work of Lekone and Finkenstädt (2006) on this epidemic provided the methodological starting point for the generalizations to stochastic SEIR modeling techniques presented in section 2.

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 term in time after that date. In order to examine the behavior of

the $\mathcal{R}_0(t)$ and $\mathcal{R}^{(EA)}(t)$ measures, 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. Here, we set their prior mean equal to these values with high precision (1000 equivalent samples). There are some other minor parameterization differences. In particular, the authors model the intensity intercept on the log scale relative to our linear predictor approach. They thus use a gamma, rather than a normal, 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. Finally, an additional run of samples was taken to estimate the two reproductive numbers.

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

The 2014 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 publication 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 a model.

Incidence data aggregated from the World Health Organization's situation report publications provides the basis for this analysis, and is modeled under overdispersion to at least 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 and visualization have been persistent features of this epidemic. The data was also lightly smoothed to ensure that a minimum of one week passed between incidence reports in order to eliminate some of the artificial variability induced by the limited availability of treatment beds, the potentially shifting geographic reach of surveillance teams, and the uneven temporal availability of incidence reports between countries.

A major choice for all epidemic models in this class is the type of intensity process to include. As there is not an established form for the intensity process in this case, we begin 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).

The next important modeling task is to choose the type of spatial structure employed.

As all three core nations share borders with each other, a natural approach would be to employ a simple neighborhood matrix. Such an approach would introduce an overall spatial correlation in the intensity process, identical between each pair of countries. It may, however, be unreasonable to assume each border is comparable in this region. Thus, 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 end up trading 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.

5. Results

5.1 *Case Study: 1995 Ebola Outbreak in Kikwit*

After running the MCMC samplers to convergence and verifying that our methods give comparable parameter estimates to the work of Lekone and Finkenstädt (2006), we examine the reproductive number estimates. In Figure 1, which corresponds to the canonical intervention specification, note that both measures follow a similar trajectory. The two curves are differentiated mainly by the smoothness of the empirically adjusted measure. Also note that the induced drop in intensity occurs slightly earlier for the empirically adjusted method due to the smoothing effect. No comparable graphic is available in Lekone and Finkenstädt (2006), although the corresponding $\mathcal{R}_0$ estimates are similar and well within the credible interval. The "effective $\mathcal{R}_0$" of Chowell et al. (2004) is not shown, being entirely indistinguishable from $\mathcal{R}_0(t)$ as the susceptible proportion is approximately 1 throughout the epidemic.

More interestingly, consider Figure 2 which demonstrates 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 (ie, an intercept). In fact, in this case $\mathcal{R}_0(t)$ and $\mathcal{R}_0$ are the same quantity. 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), which does not overlap at all with the credible interval shown here.

In contrast to the constant $\mathcal{R}_0(t)$ graph, $\mathcal{R}^{(EA)}(t)$ here distinguishes itself by shrinking towards zero after the intervention date, even though no intervention information was incorporated into the model. Of course, the $\mathcal{R}^{(EA)}$ 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 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*

5.2.1 *Model Fitting.* Examining the new case counts in Figure 3, we see that the intervention effect is not so clear as in the earlier epidemic. Cases do appear to be decreasing, but there is a lot of variability. This drives home the point that we do not have a thorough understanding of the underlying intensity process, and we therefore compare several candidate models of varying flexibility.

It is obvious from Figure 4 that the estimates for reproductive numbers in the simple inter-

cept case are underestimates. This is probably due to the requirement that these estimates reflect average behavior over a clearly heterogeneous epidemic. This heterogeneity includes a slow start in Guinea, catastrophic spread in Liberia and Sierra Leone, and any potential successes in later months. Note also that there is a lot of variability in the $\mathcal{R}^{(EA)}(t)$ which is not reflected in $\mathcal{R}_0$.

When the flexibility of the intensity process is increased by adding a three degree of freedom temporal basis (Figure 5), however, we see broad similarity in the shapes of the two reproductive number estimates and more reasonable values of $\mathcal{R}^{(EA)}(t)$. While this is a definite improvement, there is still small scale variability in $\mathcal{R}^{(EA)}(t)$ which is not captured by $\mathcal{R}_0(t)$, and the early epidemic in Sierra Leone appears to be anomalous. Moreover, the $\mathcal{R}_0(t)$ measures for Guinea and Liberia still appear to be lower than would be expected (i.e. below 1.5), but not so severely as in the intercept model.

The five degree of freedom model shown in Figure 6 is, in our judgment, broadly similar to the three degree of freedom example. The $\mathcal{R}_0(t)$ estimates are slightly more nuanced, but retain their general shape. The $\mathcal{R}^{(EA)}(t)$ estimates are quite similar. Upon considering yet more complicated temporal structures, we encountered strong evidence of overspecification, including convergence problems and excessive variability in reproductive number estimates. In light of these experiments, we chose the three degree of freedom spline basis approach as our final model. This decision was made in the full knowledge that such a basis is unlikely to completely capture temporal heterogeneity, but in the absence of structural information it appears to do a reasonable job. Moreover, given the noisy and incomplete nature of the data, much of the observed short scale variability in infection counts is likely to be spurious; overfitting is therefore an important concern.

5.2.2 *Final Model.* 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 by Liberia. Guinea had a substantially lower estimated intensity intercept. Interpretation of the three temporal basis parameters is not so straightforward, though only two 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. This view of the predictive capabilities of splines highlights a potential drawback of the method: absent a theoretical structure, the extrapolative behavior of a spline basis function need not share the same association with the data that the estimation portion does. Interpretation of credible regions for predicted quantities is therefore complicated by the fact that we have not quantified our considerable uncertainty about the way in which the intensity process should continue to evolve past the last observed data point.

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 both borders with Sierra Leone had higher estimated contact rates than the border between Guinea and Liberia. Given this observed heterogeneity, future research might explore an asymmetrical contact specification to examine the effect of any net population flows between countries.

Finally, the exposed to infectious and infectious to removed processes can be described in several ways. Quantiles are given in the supplemental material on the original parameter scale, on a daily transition probability scale, and in the context of the number of geometrically distributed compartment membership days. While not completely unreasonable, the results on this final scale emphasize the desire expressed by many authors to relax the population averaging approach to better match actual disease behavior (Porter and Oleson, 2013).

Of great interest when fitting these models is the corresponding epidemic prediction -

how long, assuming that the specified intensity process and the implied extrapolation are reasonable, is the epidemic expected to continue? For this class of models, prediction is quite straightforward. We may, for each converged MCMC sample, run the simulation forward in time for as long as desired. 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 7 are the predictions for the final selected model. Confidence bands have been omitted for clarity, and because they do not reflect the aforementioned uncertainty (i.e., they are too narrow). Interpretation of the results must be cautious. In this case, we consider the result to be weak evidence that the epidemic will continue to slow until it is contained in April or May of 2015. We observed a similar decrease in epidemic size for all fitted models, which bolsters our confidence in the prediction at least for the near future. The various models did differ significantly in the rate of decline, however, with the intercept model predicting the slowest decrease.

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 counfounded 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.

Even so, 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. We have employed graphical summary in both cases, however such an approach is neither particularly concise nor very amenable to helpful rules of thumb. In particular, requiring modelers to visually assess the similarity between reproductive number curves is both subjective and error prone. Nevertheless, the utility of this method is clear, and opportunities for further refinement remain promising.

Supplementary Materials

Web Appendices 1.1 ,1.2, and 2, referenced in Sections 2, 3, 5 respectively, are available with this paper at the Biometrics website on Wiley Online Library.

References

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

Banerjee, S. et al. (2004). *Heirarchical Modeling and Analysis for Spatial Data*. Chapman & Hall.

Brown, G. D. (2014). *spatialSEIR: An framework for fitting Bayesian spatial SEIR epidemic models using OpenCL.* R package version 0.01.

Cauchemez, S. et al. (2004). A bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statistics in Medicine* .

Cauchemez, S. et al. (2009). Household transmission of 2009 pandemic influenza a (H1N1) virus in the United States. *New England Journal of Medicine.* .

Chowell, G. et al. (2004). The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *Journal of Theoretical Biology* .

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

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

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

Frieden, T. R. et al. (2014). Ebola 2014 - new challenges, new global response and responsibility. *New England Journal of Medicine* .

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

Graves, T. L. et al. (2011). Improved mixing in MCMC algorithms for linear models. *Journal of Computational and Graphical Statistics* .

Heffernan, J. et al. (2005). Perspectives on the basic reproductive ratio. *Journal of the Royal Society. Interface* .

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

Hooten, M. B. et al. (2011). Assessing North American influenza dynamics with a statistical SIRS model. *Spatial and Spatiotemporal Epidemiology* .

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* .

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

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. (2013). A path-specific SEIR model for use with general latent and infectious time distributions. *Biometrics* **69,** 101–108.

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* .

van Boven, M. et al. (2010). Transmission of novel influenza a(h1n1) in households with post-exposure antiviral prophylaxis. *PLoS One* **5,**.

Verdasca, J. et al. (2005). Recurrent epidemics in small world networks. *Journal of Theoretical Biology* .

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

Figure 1. 1995 Ebola Outbreak in Kikwit: SEIR Model With Intervention Parameter

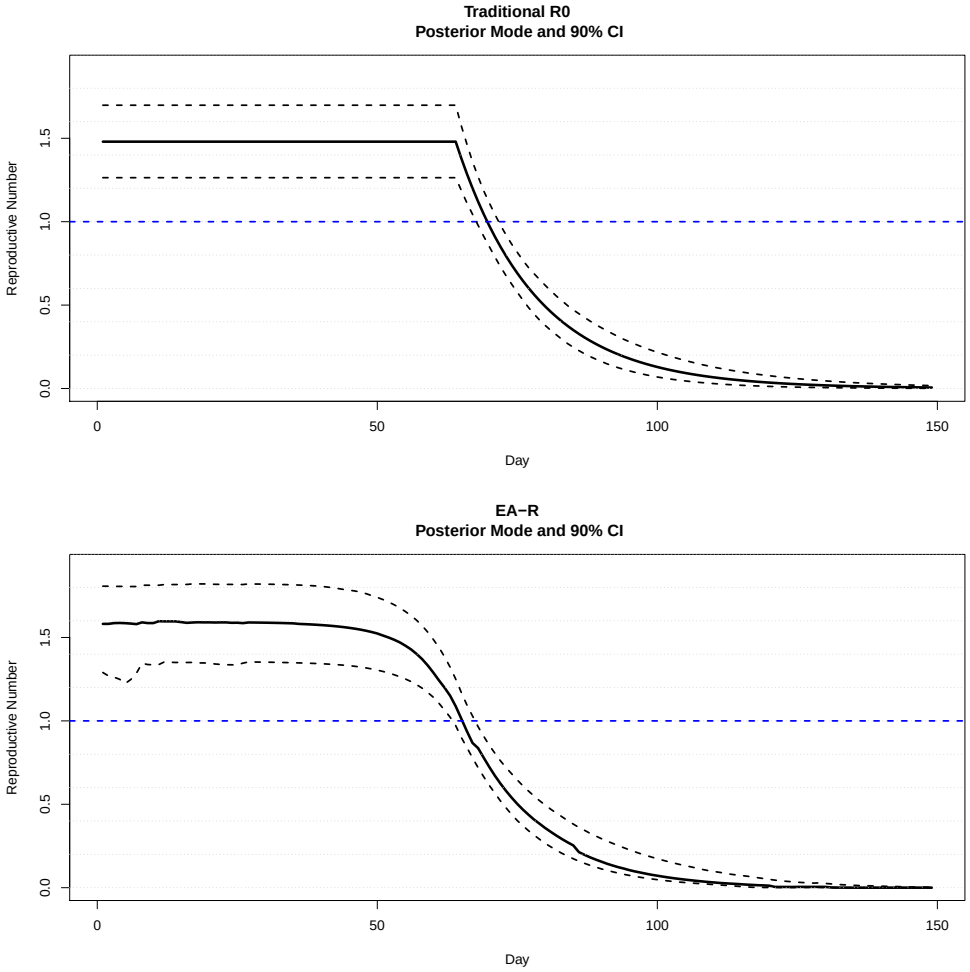

Figure 2. 1995 Ebola Outbreak in Kikwit: Single Parameter SEIR Model

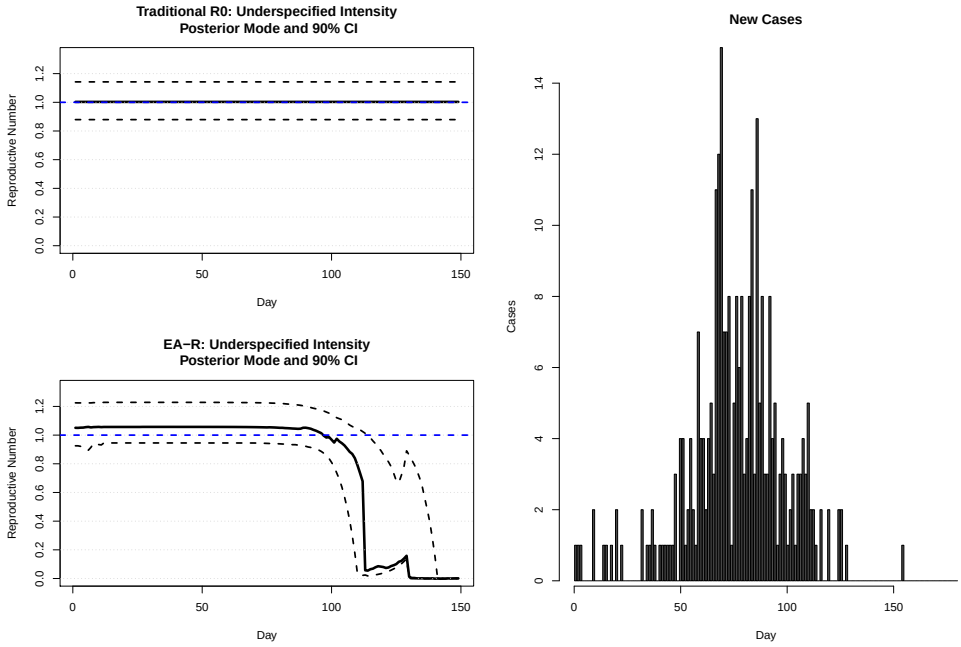

Figure 3. 2014 Ebola Outbreak in West Africa: Estimated Infections Per Day

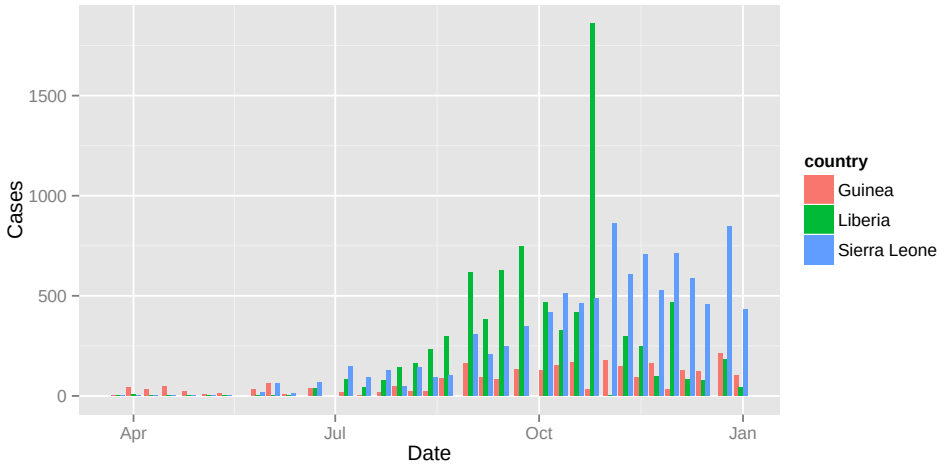

Figure 4. 2014 Ebola Outbreak in West Africa: Country Intercept Model with HPD Estimate and 90% CI for $\mathcal{R}_0$ and $\mathcal{R}^{(EA)}(t)$.

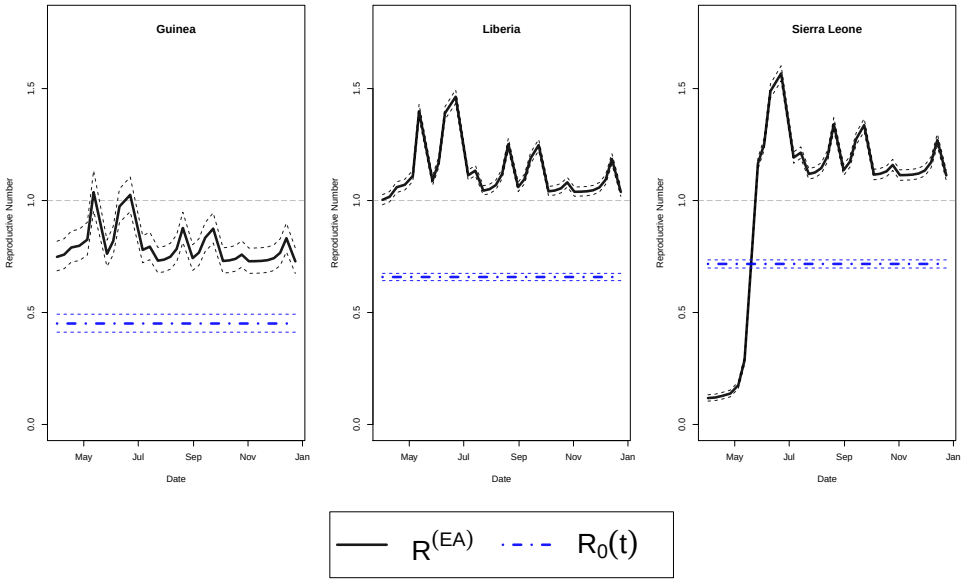

Figure 5. 2014 Ebola Outbreak in West Africa: Three Degree of Freedom Basis Model with HPD Estimate and 90% CI for $\mathcal{R}_0(t)$ and $\mathcal{R}^{(EA)}(t)$.

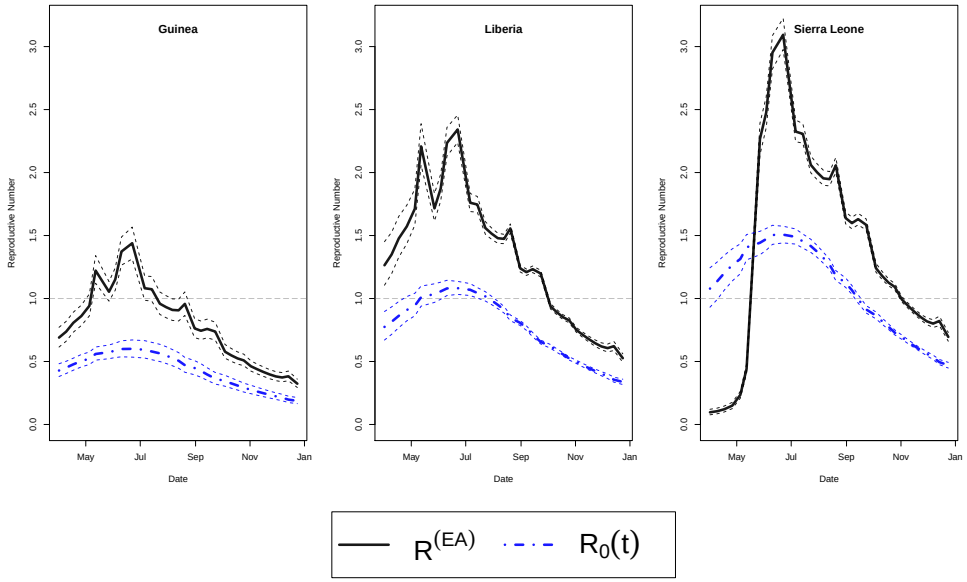

Figure 6. 2014 Ebola Outbreak in West Africa: Five Degree of Freedom Basis Model with HPD Estimate and 90% CI for $\mathcal{R}_0(t)$ and $\mathcal{R}^{(EA)}(t)$.

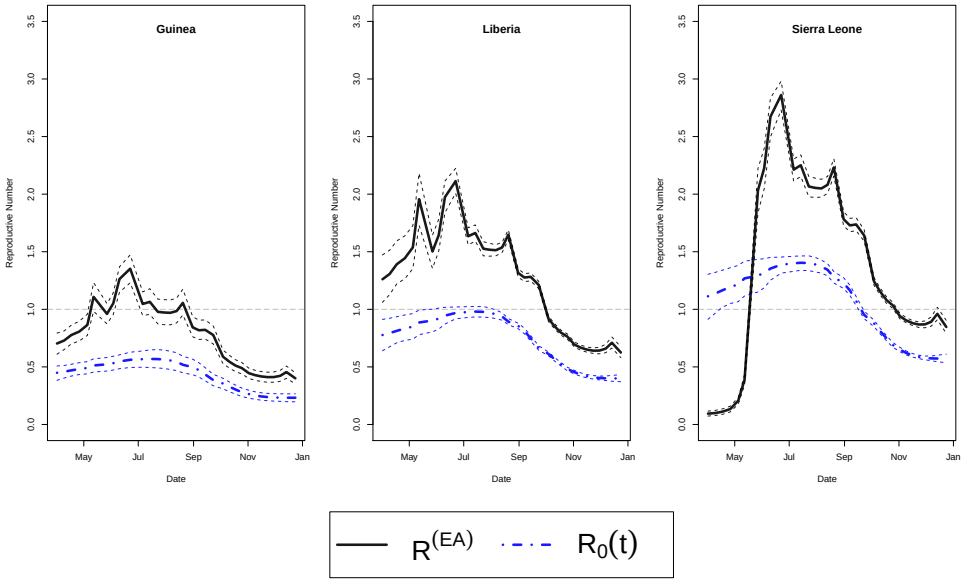

Figure 7. Long Term Infectious Count Predictions for the 2014 Outbreak: Three Degree of Freedom Model