Biometrics 000, 000–000 DOI: 000
000 0000

Web Based Supplementary Materials for 'An Empirically Adjusted Approach to Reproductive Number Estimation for Stochastic Compartmental Models'

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

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

1. Important Derivations

1.1 *Exposure Probability: Parametric Form*

Define δ_{ij} to be the proportion of persons who are infectious in spatial unit s_j at time t_i. Then, letting $\text{Inf}(s_j, t_i)$ denote the event that a person becomes infected from contact within spatial unit s_j at time t_i and $!\text{Inf}(s_j, t_i)$ denote its complement. We may then partition the infection probability as in (1).

$$P(\text{Inf}(., t_i)) = 1 - P(!Inf(s_j, t_i)) \cdot P(!Inf(s_{-j}, t_i)) \tag{1}$$

where

$$\begin{aligned} P(!\text{Inf}(s_j, t_i)) = E(!\text{Inf}(s_j, t_i)) &= E(E(!\text{Inf}(s_j, t_i)|K = k)) \\ &= E((1 - \delta_{ij}p)^k) \\ &= \sum_{k=0}^{\infty}(1 - \delta_{ij}p)^k(\frac{\lambda_{ij}^k e^{-\lambda_{ij}}}{k!}) \\ &= \sum_{k=0}^{\infty} q_{ij}^k(\frac{\lambda_{ij}^k e^{-\lambda_{ij}}}{k!}) \\ = \frac{e^{-\lambda_{ij}}}{e^{-q_{ij}\lambda_{ij}}}(1) = e^{-\lambda_i \cdot (1-q_{ij})} &= e^{-\lambda_i \cdot p_{ij}} = e^{-\lambda_i \cdot (\delta_{ij}p)} \end{aligned}$$

Therefore,

$$P(\text{Inf}(s_j, t_i)) = 1 - e^{-\lambda_{ij} \cdot (\delta_{ij}p)}$$

Similarly,

$$P(!\text{Inf}(s_{-j}, t_i)) = \prod_{\{l \neq j\}} [P(!\text{Inf}(s_l, t_i))]$$

$$= \prod_{\{l \neq j\}} [E(!\text{Inf}(s_{-j}, t_i))] = \prod_{\{l \neq j\}} [E(E(!\text{Inf}(s_{-j}, t_i)|K = k))]$$

$$= \prod_{\{l \neq j\}} \left[E(1 - \delta_{il}p)^k)\right]$$

$$= \prod_{\{l \neq j\}} \left[\sum_{k=0}^{\infty} (q_{il}(j))^k \frac{(\lambda_{il} \cdot f(d_{jl}))^k e^{-\lambda_{il} \cdot f(d_{jl})}}{k!}\right] = \prod_{\{l \neq j\}} \left[\frac{e^{-\lambda_{il} \cdot f(d_{jl})}}{e^{-q_{il}\lambda_{il} f(d_{jl})}}(1)\right]$$

$$= \prod_{\{l \neq j\}} \left[e^{-\lambda_{il} \cdot f(d_{jl}) p_{il}}\right] = \prod_{\{l \neq j\}} \left[e^{-\lambda_{il} \cdot f(d_{jl}) \cdot (\delta_{il} p)}\right]$$

$$= \exp\left\{\sum_{\{l \neq j\}} [p\lambda_l \delta_{il} f(d_{jl})]\right\}$$

Thus, for the probability of infection for a person living in s_j at time t_i we have:

$$1 - \left(e^{-\lambda_{ij} \cdot (\delta_{ij} p)}\right) \left(e^{\left\{\sum_{\{l \neq j\}} \left[p\lambda_{il}\delta_{il} f(d_{jl})\right]\right\}}\right)$$

$$= 1 - \exp\left\{-\delta_{ij} e^{\theta_{ij}} - \sum_{\{l \neq j\}} (f(d_{jl})\delta_{il} e^{\theta_{il}})\right\}$$

where $\theta_{ij} = \log(\lambda_{ij} p)$

1.2 $\mathcal{R}_0$ *as a special case of* $\mathcal{R}^{(\mathcal{EA})}$

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, $S_t \approx N \; \forall \; t$. Further, assume a simple single parameter intensity function with 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:

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

Note that, taking the limit as $(N \to \infty)$, this first term is indeterminate: $\frac{0}{0}$. Thus, with an application of L'Hospital's rule we have the limit expression

$$\begin{aligned}&\lim_{N\to\infty} \frac{-\frac{1}{N^2}e^{\theta}\exp\{-\frac{1}{N}e^{\theta}\}}{-\frac{1}{N^2}}\left[\sum_{t=t_i}^{t_\infty}(1-\pi^{(IR)})^{(t-t_i)}\right]. \\ &= e^{\theta}\left[\sum_{t=t_i}^{t_\infty}(1-\pi^{(IR)})^{(t-t_i)}\right].\end{aligned}$$

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 $(N \to \infty)$, by the convergence property of the geometric series we have

$$\frac{e^{\theta}}{1-(1-\pi^{(IR)})} = \frac{e^{\theta}}{\pi^{(IR)}}$$

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.

2. Additional statistical output

2.1 *Simulations*

As mentioned in our primary manuscript, considerable attention to detail is required to examine the behavior of the empirically adjusted reproductive number under simulation. This quantity describes the expected number of secondary infections caused over the temporal course and spatial domain of an epidemic; if we consider only the spatial SEIR model class as parameterized here, insufficient information is available to calculate the true number of such secondary infections even when all model parameters are known, though we can compute its expectation. Therefore, we begin by simulating data on an individual level, from a so called "agent based" perspective, according to an epidemic process which is analogous to the spatial SEIR framework defined for individuals rather than aggregate spatial units. We then consider traditional population averaged simulations, and examine the ability of the empirically adjusted reproductive number to both estimate the true expected number of secondary infections, as well as approximate and be profitably compared to traditional reproductive numbers under various forms of misspecification.

In each of the following simulations, the intensity process was assumed to depend on a baseline parameter and a linear intervention effect, similar to the analysis of Ebola data from Kikwit.

2.1.1 *Agent Based Simulation.* In the agent based setting, we simulate 100 epidemics in which a fixed population of 500 susceptible, 10 infectious, and 10 exposed individuals are tracked over 150 time points. Agents were considered homogeneous, and contact probabilities were equal for each pair of agents. Contact events which resulted in a new infection were recorded, along with the current compartmental states of the agents. This individual level tracking enabled us to calculate the true average number of secondary infections over time. Once the data was simulated, it was aggregated by a temporal index to provide a vector of new infection counts for analysis. Models were fitted using both the true intensity process and an underspecified version with only an intercept, and bias was computed over time for the three reproductive numbers.

As briefly noted in Section 4, $\mathcal{R}^{(EA)}$ provides a uniformly better estimate of the true number of secondary infections over time in a particular epidemic than that provided by $\mathcal{R}_0(t)$. Interestingly, in this relatively small population, the effective reproductive number does a comparable job of capturing the expected number of secondary infections, with the exception of the temporal region in which the effect of intensity process underspecification is most pronounced: the intervention period. During this portion of the simulated epidemics, the empirically adjusted measure begins to shrink towards the true value as model misspecification grows more severe. For well specified models in this setting, the absolute bias of $\mathcal{R}^{(EA)}$ was less than that of $\mathcal{R}_0(t)$ for an average of 70.6% of time points, and less than that of effective $\mathcal{R}_0(t)$ for an average of 51.7%. In underspecified analyses, $\mathcal{R}^{(EA)}$ similarly outperformed these two measures by 73.9% and 52.6%, respectively.

2.1.2 *Population Averaged Simulation - Single Location.* We next simulate 100 epidemics directly from the population averaged epidemic process described in Section 2. A large population of susceptibles (5.4M) and a single infectious individual, modeled on the population of Kikwit, were tracked over 150 time points. We again consider correctly and underspecified intensity processes.

The results were mixed in this setting. While we expect $\mathcal{R}^{(EA)}(t)$to share many properties with $\mathcal{R}_0(t)$, it was not designed to estimate the same quantity. Nevertheless, for underspecified intensity processes, $\mathcal{R}^{(EA)}$ maintained a lower absolute distance from the true time varying basic reproductive number an average of 71.3% of epidemic time. Unsurprisingly, when the model was correctly specified, this advantage dropped to 14.7%. When we consider the ability of these quantities to estimate the true expected number of secondary infections, however, $\mathcal{R}^{(EA)}$ uniformly outperforms the other measures.

As with the bias, the $\mathcal{R}^{(EA)}$ estimate of $\mathcal{R}_0(t)$ thresholding regions (i.e., a comparison to 1.0) outperforms the traditional method in the presence of misspecification (66.4% vs. 50%), loses under the correct specification (91.3% vs. 98.7%), and uniformly outperforms $\mathcal{R}_0(t)$ when estimating the true expectation. Finally, we note that the average correlation between $\mathcal{R}^{(EA)}(t)$ and effective $\mathcal{R}_0(t)$ was 0.92 for well specified models, and 0.45 for underspecified models. This observation agrees with our informal observation that these measures deviate more in the presence of model misspecification.

Additional results are available in Tables 1 and 2. Note that $\mathcal{R}_0(t)$ and effective $\mathcal{R}_0(t)$ are mathematically and conceptually interchangeable in this setting, as the high susceptible fraction diminishes any scaling effects.

2.1.3 *Population Averaged Simulation - Spatial Setting.* We now consider the important case of spatial misspecification. Due to the difficulty of obtaining detailed structural information about populations, we are often obliged to assume that large groups of individuals

are homogeneous with respect to susceptibility and mixing behavior. This is, of course, very unlikely to be true. In order to examine the effect of spatial misspecification on our reproductive number estimates, we define a set of three spatial locations: $\{S_1, s_2, s_3\}$ with population sizes of 2.2M, 1M, and 2.2M, respectively, and a single infectious individual in each unit. We further posit the existence of only two borders: $\{s_1, s_2\}$ and $\{s_2, s_3\}$. Models were fit using the correct spatial structure and an underspecified version which introduces only a single spatial autocorrelation parameter, neglecting the heterogeneity between location borders. Results are again presented in Tables 1 and 2, and demonstrate that traditional reproductive measures differ from the empirically adjusted formulation more dramatically in spatially heterogeneous settings. While $\mathcal{R}^{(EA)}$continues to outperform $\mathcal{R}_0(t)$ when estimating the population specific expected number of secondary infections, in the spatial setting it does not do a particularly good job of approximating the traditional time varying basic reproductive number. Examination of the reproductive number estimates reveals that this particularly reflects large differences between the basic and adjusted reproductive numbers early in the epidemic. Only one of three spatial locations initially contains infectious individuals, which gives rise to large differences between the empirically adjusted and traditional measures; only the empirically adjusted measure depends on the number of infectious individuals.

[Figure 1 about here.]

[Figure 2 about here.]

[Table 1 about here.]

[Table 2 about here.]

2.2 *West Africa Ebola Analysis*

Here we compare the empirically adjusted and basic reproductive number curves for intensity process models of varying flexibility. Recall that, for large populations and moderately sized populations, the effective reproductive number is approximately equal to $\mathcal{R}_0(t)$.

It is obvious from Figure 3 that the estimates for reproductive numbers in the simple intercept 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 $\mathcal{R}^{(EA)}(t)$ which is not reflected in $\mathcal{R}_0$.

When the flexibility of the intensity process is increased by incorporating a three degree of freedom temporal spline basis in addition to the intercept (Figure 4), 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 results from the five degree of freedom model shown in Figure 5 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. Given the absence of

structural information, however, this simple model 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.

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Table 3 about here.]

[Figure 6 about here.]

2.3 *Updated West Africa Results - Predictive Accuracy*

One challenge to the analysis of ongoing epidemics is that the situation evolves much more rapidly than even the most efficient academic journal. With that in mind, our originally submitted predictions were made in early January 2015. All such predictions must be interpreted cautiously, especially when the results might be used by policymakers in decisions which affect the wellbeing of the population under study. Even so, an updated analysis performed in April 2015 demonstrates the reasonableness of the previous predictions. While different dynamics appear to have dominated once the estimated epidemic size began to drop below around 250, we did indeed observe a continued decrease. This was particularly true of Sierra Leone. Presented in Figure 7 is the incidence data on which the updated analysis was based, while Figure 8 shows the resulting estimated and predicted epidemic size curves.

[Figure 7 about here.]

[Figure 8 about here.]

Figure 1. Population Averaged Single Location Simulation

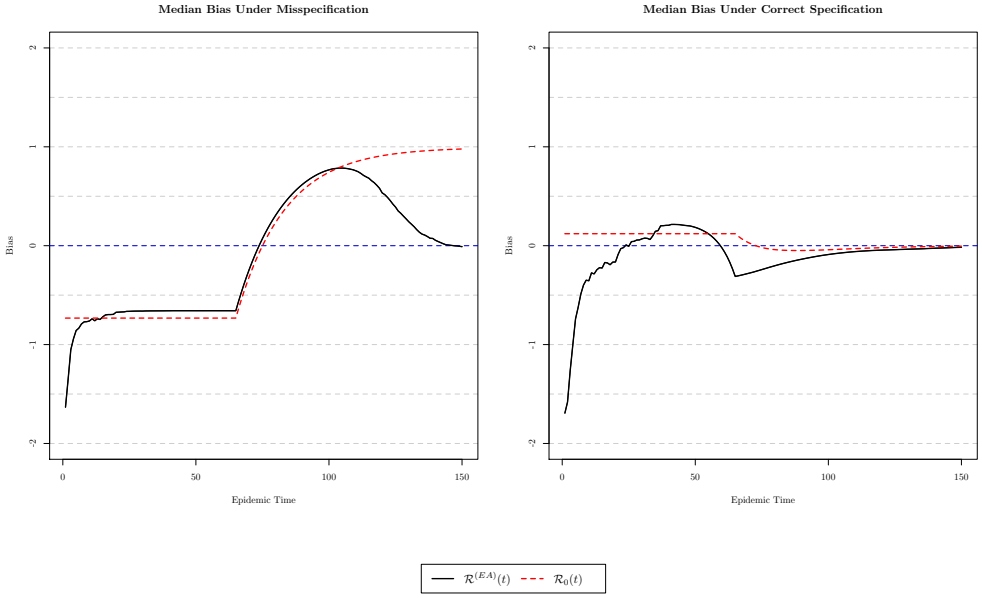

Figure 2. Agent Based Single Location Simulation

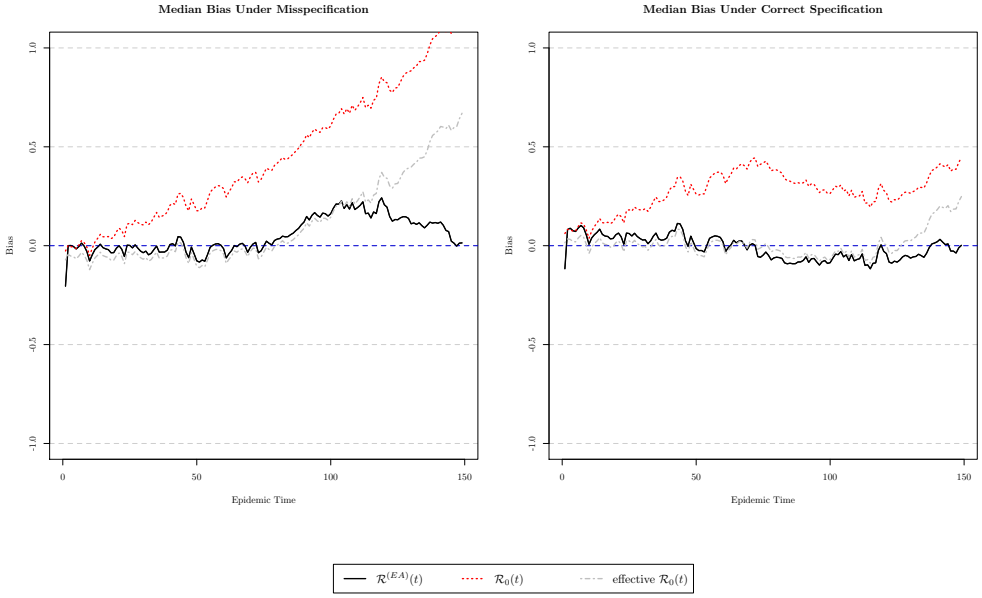

Figure 3. 2014 Ebola Outbreak in West Africa: Country Intercept Model with HPD Estimate and 90% CI for $\mathcal{R}_0$ and $\mathcal{R}^{(EA)}(t)$.

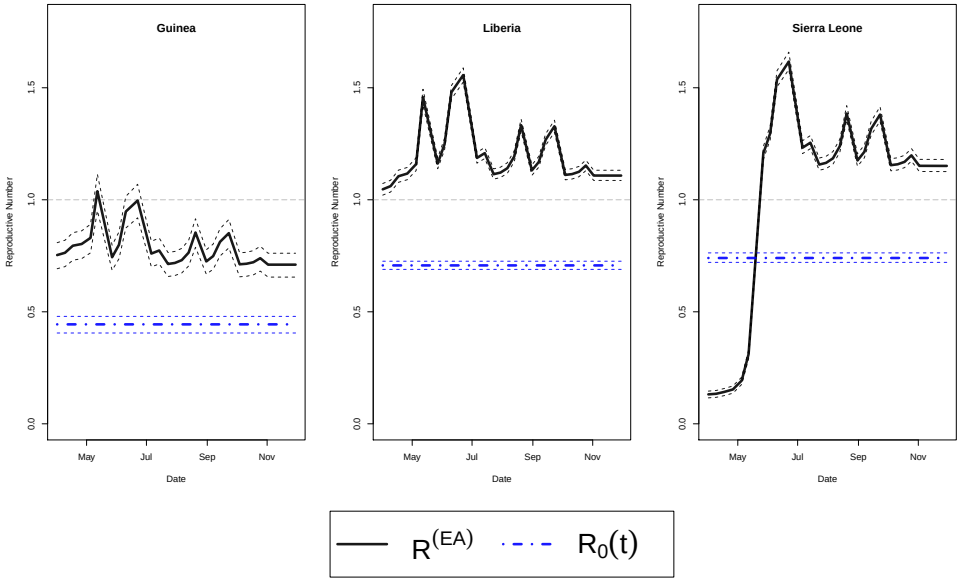

Figure 4. 2014 Ebola Outbreak in West Africa: Three Degree of Freedom Basis Model with HPD Estimate and 90% CI for $\mathcal{R}_0(t)$ and $\mathcal{R}^{(EA)}(t)$.

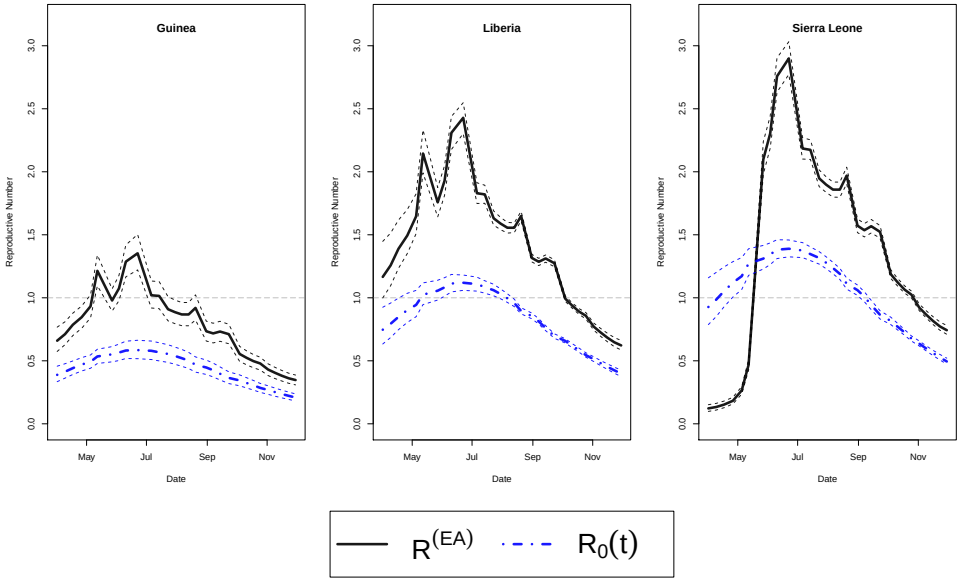

Figure 5. 2014 Ebola Outbreak in West Africa: Five Degree of Freedom Basis Model with HPD Estimate and 90% CI for $\mathcal{R}_0(t)$ and $\mathcal{R}^{(EA)}(t)$.

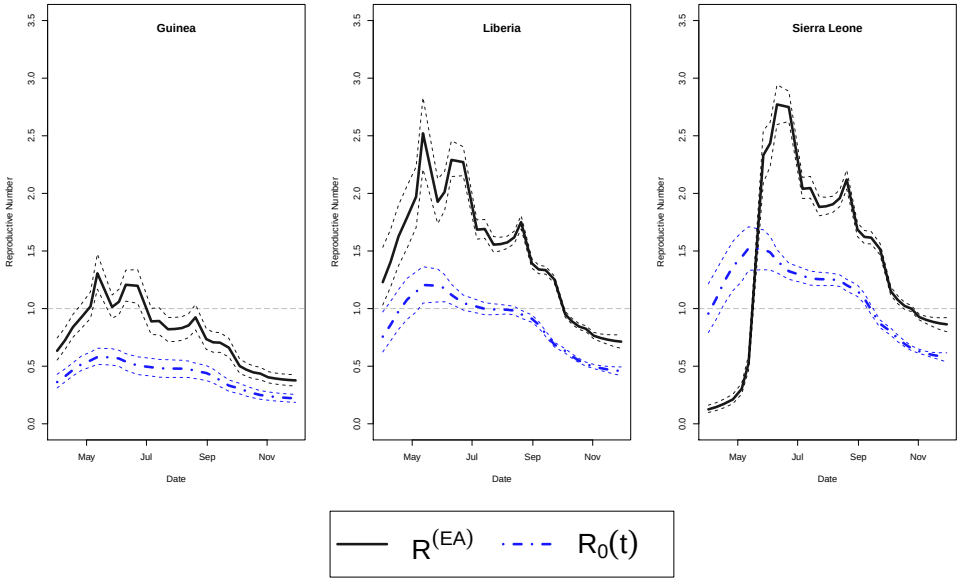

Figure 6. Three Degree of Freedom Spline Basis for Estimation and Prediction

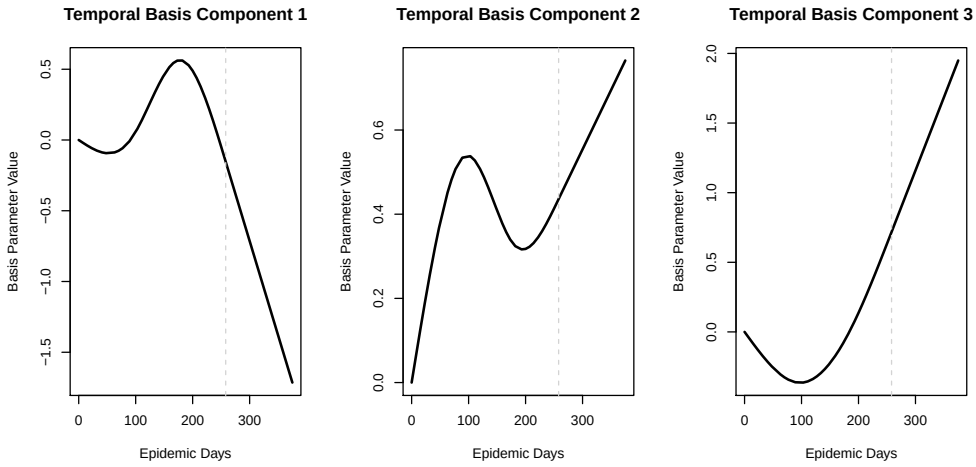

Figure 7. Updated Predictions - April 2015

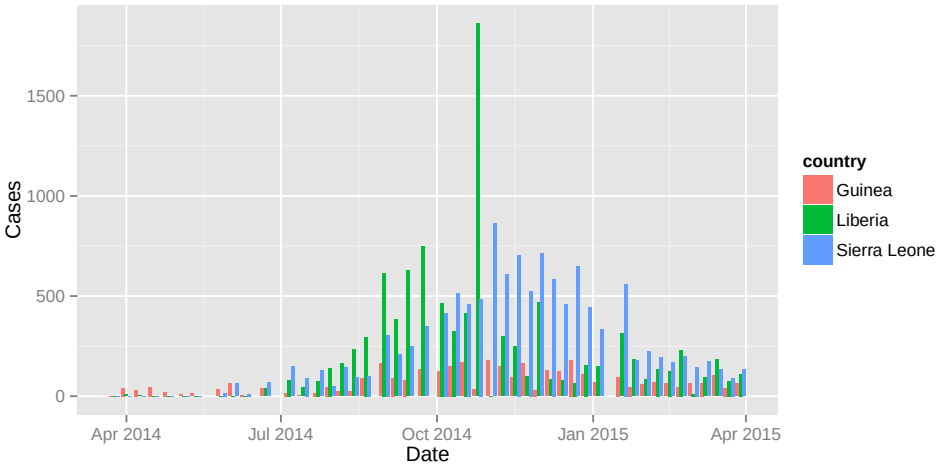

Figure 8. Updated Predictions - April 2015

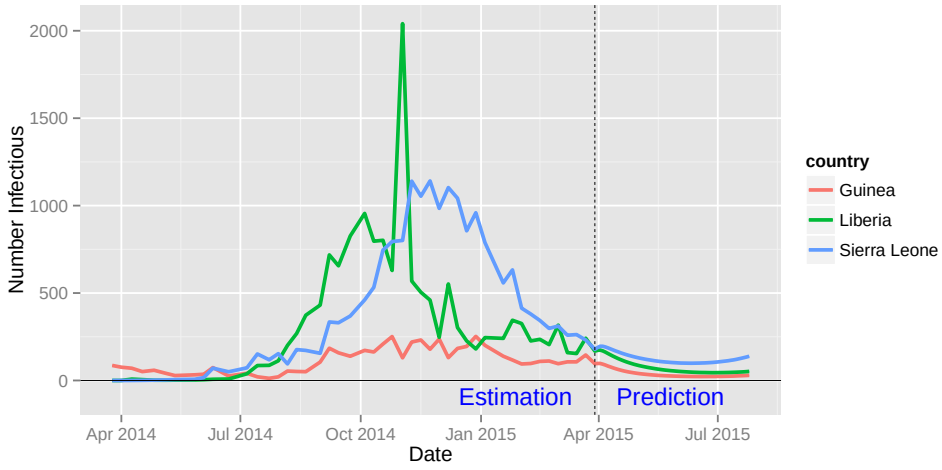

Table 1
Population Averaged Simulations - Concordance with Reproductive Number Thresholds

Simulation	Specification	Estimator	$\mathcal{R}_0$ Concordance	$\mathcal{R}^{(EA)}$Concordance
Simple	Under	Eff. $\mathcal{R}_0(t)$	0.500	0.502
Simple	Under	$\mathcal{R}^{(EA)}(t)$	0.664	0.651
Simple	Full	Eff. $\mathcal{R}_0(t)$	0.987	0.963
Simple	Full	$\mathcal{R}^{(EA)}(t)$	0.912	0.926
Spatial	Under	Eff. $\mathcal{R}_0(t)$	0.995	0.856
Spatial	Under	$\mathcal{R}^{(EA)}(t)$	0.857	0.959
Spatial	Full	Eff. $\mathcal{R}_0(t)$	0.996	0.859
Spatial	Full	$\mathcal{R}^{(EA)}(t)$	0.844	0.952

Table 2
Population Averaged Simulations - Proportion Exhibiting Minimum Absolute Bias

Simulation	Specification	Estimator	Estimating $\mathcal{R}_0(t)$	Estimating $\mathcal{R}^{(EA)}(t)$
Simple	Under	Eff. $\mathcal{R}_0(t)$	0.287	0.313
Simple	Under	$\mathcal{R}^{(EA)}(t)$	0.713	0.687
Simple	Full	Eff. $\mathcal{R}_0(t)$	0.853	0.640
Simple	Full	$\mathcal{R}^{(EA)}(t)$	0.147	0.360
Spatial	Under	Eff. $\mathcal{R}_0(t)$	0.847	0.167
Spatial	Under	$\mathcal{R}^{(EA)}(t)$	0.153	0.833
Spatial	Full	Eff. $\mathcal{R}_0(t)$	0.887	0.261
Spatial	Full	$\mathcal{R}^{(EA)}(t)$	0.113	0.739

Table 3
MCMC Quantiles for Three Degree of Freedom Model Parameters

	2.5%	25%	50%	75%	97.5%
Guinea Intercept	-3.26	-3.11	-3.03	-2.95	-2.80
Liberia Intercept	-2.67	-2.49	-2.38	-2.28	-2.07
Sierra Leone Intercept	-2.45	-2.28	-2.17	-2.06	-1.85
Time component 1	-0.27	-0.13	-0.06	0.02	0.15
Time component 2	-0.54	-0.12	0.09	0.31	0.65
Time component 3	-1.10	-1.02	-0.98	-0.93	-0.85
Spatial Dependence Parameter 1	0.02	0.03	0.03	0.04	0.04
Spatial Dependence Parameter 2	0.02	0.03	0.04	0.04	0.05
Spatial Dependence Parameter 3	0.06	0.08	0.08	0.09	0.10
Overdispersion Precision	6.54	6.62	6.67	6.71	6.80
E to I Transition Parameter	0.19	0.19	0.19	0.19	0.20
I to R Transition Parameter	0.13	0.14	0.14	0.14	0.14
E to I Transition Probability	0.17	0.17	0.18	0.18	0.18
I to R Transition Probability	0.13	0.13	0.13	0.13	0.13
Days in Exposed Category	0.00	1.00	3.00	7.00	18.00
Days in Infectious Category	0.00	2.00	5.00	10.00	26.00