New model diagnostics for spatio-temporal systems in epidemiology and ecology

Lau, Max Siu Yin; Marion, Glenn; Streftaris, George; Gibson, Gavin Jarvis

Published in:
Journal of the Royal Society Interface

DOI:
10.1098/rsif.2013.1093

Publication date:
2014

New model diagnostics for spatio-temporal systems in epidemiology and ecology

Max S.Y. Lau[1,3], Glenn Marion[3], George Streftaris[1,2], and Gavin J. Gibson[1,2]

[1]Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, United Kingdom, EH14 4AS
[2]Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, United Kingdom, EH14 4AS
[3]Biomathematics and Statistics Scotland, Edinburgh, United Kingdom, EH9 3JZ

Abstract

A cardinal challenge in epidemiological and ecological modelling is to develop effective and easily deployed tools for model assessment. The availability of such methods would greatly improve understanding, prediction and management of disease and ecosystems. Conventional Bayesian model assessment tools such as Bayes factors and the Deviance Information Criterion (DIC) are natural candidates but suffer from important limitations due to their sensitivity and complexity. Posterior predictive checks, which utilize summary statistics of the observed process simulated from competing models, can provide a measure of model fit but appropriate statistics can be difficult to identify. Here we develop a novel approach for diagnosing mis-specifications of a general spatio-temporal transmission model by embedding classical ideas within a Bayesian analysis. Specifically, by proposing suitably designed non-centered parameterization schemes, we construct latent residuals whose sampling properties are known given the model specification and which can be used to measure overall fit and to elicit evidence of the nature of mis-specifications of spatial and temporal processes included in the model. This model assessment approach can readily be implemented as an addendum to standard estimation algorithms for sampling from the posterior distributions such as Markov chain Monte Carlo. The proposed methodology is first tested using simulated data and subsequently applied to data describing the spread of Heracleum mantegazzianum (giant hogweed) across Great Britain over a thirty-year period. The proposed methods are compared with alternative techniques including posterior predictive checking and the DIC. Results show that the proposed diagnostic tools are effective in assessing competing stochastic spatio-temporal transmission models and may offer improvements in power to detect model mis-specifications. Moreover, the latent-residual framework introduced here extends readily to a broad range of ecological and epidemiological models.

Keywords: Spatio-temporal model assessment | Latent residuals | Non-centered parameterization | Bayesian Inference

1 Introduction

Stochastic spatio-temporal models are playing an increasingly important role in epidemiological and ecological studies relating to transmission of diseases [1, 2], invasion of alien species [3] and population movements in response to climate changes [4]. It is well known that the predicted dynamics of such systems can be extremely sensitive to the choice of model, with consequent implications for the design of control strategies [1, 5, 6, 7], but as yet there is a lack of effective model assessment tools described in the literature. For example, studies of foot and mouth disease have cited the importance of selecting between a long-tailed spatial kernel (see Section 2) versus a localized spatial kernel, but this model-choice problem is far from being resolved [8, 5]. Further model-choice problems arise in relation to the parametric form of the distributions of incubation and infectious periods in models of measles [9, 10, 11], and in relation to diseases such as smallpox [12] and AIDS [13].

Bayesian model assessment techniques appear appealing [14, 15, 16] particularly since many of the above studies use Bayesian techniques for model fitting. However it is well known that this approach is extremely sensitive to prior

Proposition 1.1 *A threshold model, which states that the k^{th} infection event would occur at time t, where $t \geq t_{k-1}$, only if $A_{k-1}(t) \geq r_{1k}$, is equivalent to the mechanism specified by equation* (1).

Proof Denoting $\mathcal{F}_t$ as the history of the change of infectiousness before time t, the probability of having the k^{th} infection event at time interval $(t; t+dt)$, where $t \geq t_{k-1}$, from the threshold model defined in proposition 1.1 is given by the following,

$$\begin{aligned}
&P(A_{k-1}(t) \leq r_{1k} \leq A_{k-1}(t) + dA_{k-1}(t) | r_{1k} > A_{k-1}(t); t_{k-1}; \mathcal{F}_t) \\
&= h_G(A_{k-1}(t)) dA_{k-1}(t) + o(dA_{k-1}(t)) \\
&= dA_{k-1}(t) + o(dt) \quad (\because h_G(\cdot) = 1) \\
&= dQ(t) + o(dt) \\
&= \sum_{j \in \ _S(t)} \{ \quad + \quad \sum_{i \in \ _I(t)} K(d_{ij}) \} dt + o(dt).
\end{aligned}$$

The second last equality holds as

$$A_{k-1}(t) = Q(t) - Q(t_{k-1}). \quad \blacksquare$$

Further, conditional on the occurrence of the k^{th} infection event at time t_k, we can construct the corresponding infection link within the competing-risk framework. We first denote p_{ij} as the probability of individual i infecting individual j where $i \in \ _I(t_k)$ and $j \in \ _S(t_k)$. By noting that $p_{ij} = z_{ij}\, dt$, where $z_{ij} =$ when considering the primary infection and $z_{ij} = K(d_{ij})$ for $i \in \ _I(t_k)$ and $j \in \ _S(t_k)$, it can be readily seen that the total probability of having this infection event is in fact the sum of all p_{ij} and this sum is the same as the transmission probability in equation (1) and hence the equivalence is not altered in constructing the infection link corresponding to this infection event. The actual infection link is then determined by a random draw from $U(0; 1)$ and the values of z_{ij} : we first sort all the $p'_{ij} = z_{ij} = \sum_{i,j} z_{ij}$ in ascending order and denote them as $p'_{(1)}, \ldots, p'_{(m)}$ where m is the total number of the possible links; we then draw a random number, r_{2k}, from $U(0; 1)$ (i.e., the ILR), if $r_{2k} \in [p'_{(n)}; p'_{(n+1)})$ n^{th} link is realized as the actual infection link.

1.3 Construction of the Sojourn Time

We now propose a threshold model for the construction of the sojourn time in class E (i.e., the latent period) and prove its equivalence with the mechanism defined in equation (2). Define the *pressure* $Q_w(t)$ exerted on an individual corresponding to the k^{th} infection event by time t by the following expression,

$$Q_w(t) = \int_0^t h_T(y) dy \tag{4}$$

where $t = 0$ corresponds to the exposure time of the individual. Also define the *threshold* r_{3k} (i.e., LTR) such that

$$r_{3k} \sim Exp(1).$$

Similarly, r_{3k} can be transformed to $U(0; 1)$ by the cumulative distribution function of the exponential distribution. Finally define $h_{\ w}(\cdot)$ to be the hazard function of the (random) threshold r_{3k}.

Proposition 1.2 *A threshold model, which states that the transition of an exposed individual k from class E to class I would occur at time t only if $Q_w(t) \geq r_{3k}$ (i.e. $t = \inf\{ t' \mid Q_w(t') \geq r_{3k} \}$), is equivalent to the mechanism specified by equation* (2).

Proof The transition probability from class E to class I during time interval $(t; t+dt)$ from the threshold model defined in proposition 1.2 is given by the following,

$$\begin{aligned}
&P(Q_w(t) < r_{3k} \leq Q_w(t) + dQ_w(t) | k \in \ _E(t); r_{3k} > Q_w(t)) \\
&= h_{\ w}(Q_w(t)) dQ_w(t) + o(dQ_w(t)) \\
&= dQ_w(t) + o(dt) \quad (\because h_{\ w}(\cdot) = 1) \\
&= h_T(t) dt + o(dt) \quad \blacksquare
\end{aligned}$$

Denote $F_{\ w}(\cdot)$ and $F_T(\cdot)$ as the cumulative distribution functions for the threshold and for the sojourn time in class E respectively. It should be noted that $F_T(t) = F_{\ w}(Q_w(t))$ where $F_{\ w}(Q_w(t)) \sim U(0; 1)$. Therefore, the sojourn time can be obtained by computing $F_T^{-1}(r')$ where $r' \sim U(0; 1)$.

It can be readily seen that this approach can be extended to transition between two classes in which the forms of sojourn time distributions are explicit (e.g., sojourn time in class I).

1.4 Rationale for Ordering the Infection Links

The ordering of infection links for an exposure event according to the strength of the links p_{ij} is not necessary when one only wishes to generate stochastic realisations of the epidemic using a functional-model representation. However the operation is crucial to ensuring that the imputed $\widetilde{r}_2$ may be informative regarding possible mis-specification of the kernel. In particular, our goal is to distinguish between the comparative goodness-of-fit of radially symmetric kernels that differ in terms of their behaviours at short and at long distances. When the ordering operation is used, the imputed residuals corresponding to infection links that represent very long or very short range transmissions are located at the extremes of the unit interval. A mis-specified kernel, which misrepresents the propensity for short- or long-range transmission, may therefore be expected to cause the distribution of imputed $\widetilde{r}_2$ to deviate from $U(0;1)$ by exhibiting a concentration, or a scarcity, of residuals at the extremes of the unit interval, dependent on the nature of the mis-specification. Our results presented in the main text suggest that this hoped-for sensitivity is indeed achieved.

To illustrate the point further we consider the case where an exponentially-bounded kernel is fitted to data generated using a power-law kernel (corresponding to Scenario I in section 2.1 of the main text). Figure S1 is a schematic representation of the relative strength of interaction of these kernels. The strength of infection links are represented by the segments length, and the monotonically decreasing nature of these kernels means that short range links are associated with stronger links than longer range interactions. Figure S1 shows the tendency of the exponential kernel to underestimate the interaction strength at short and long distances; as a result, the lengths of these infection links (which correspond to short and long distances transmission) are reduced when an exponential kernel is fitted. Should these links are imputed as the active links for exposure events, then the corresponding imputed residuals will be located closer to the extremes of the unit interval, than had the correct kernel been used, leading to a concentration of residuals at the extremes (also see Figure 2 in the main text) in this case. Note that the above ordering operation is specifically aimed at comparing radially symmetric kernels with different tail properties, a common goal in epidemic studies. Alternative ordering schemes may be considered if our prior knowledge suggested a different form of mis-specification. For example, if transmission were to occur preferentially in certain directions, for example due to prevailing wind or other effects, then an ordering operation that took account of the direction, as well as the length, of the I-S links may be advisable. Such developments are beyond the scope of this paper.

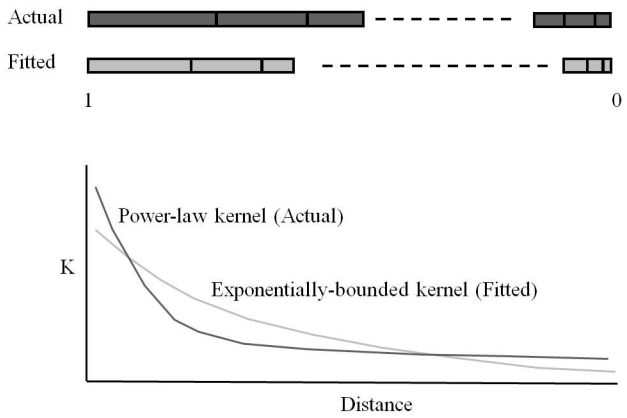

Figure S1: A schematic representation of the relative strength of interaction of these kernels. The strength of infection links are represented by the segments length.

2 Statistical Inference

2.1 Likelihood

We consider an epidemic from a spatio-temporal *S-E-I-R* model. Consider a population with size N. Assume that individuals in the population are all susceptible at time 0 which is the time of introduction of the force of primary infection, and assume that the epidemic is to be observed up to time $t_{\max}$. Let $\mathbf{E} = (E_1; E_2; \ldots; E_{N_E})$ be a vector of the exposure times of N_E individuals, $\mathbf{I} = (I_1; I_2; \ldots; I_{N_I})$ be a vector of the times of becoming infectious of N_I individuals; $\mathbf{R} = (R_1; R_2; \ldots; R_{N_R})$ be a vector of the times of recovery of N_R individuals. Also,

we let $_{\mathrm{U}}$ be the set of indices of the individuals remaining in class S at the end of observation period $t_{\max}$; and, we let $_{\mathrm{E}}$, $_{\mathrm{I}}$ and $_{\mathrm{R}}$ be the set of indices of the individuals who have gone through class E, class I and class R by $t_{\max}$ respectively. Also, we let $_{\mathrm{E}\setminus\mathrm{I}}$ be the set corresponding to the exposed individuals who have not been infectious up to time $t_{\max}$ and $_{\mathrm{I}\setminus\mathrm{R}}$ be the set corresponding to the infectious individuals who have not recovered up to time $t_{\max}$. Similar to previous sections, we let be the primary infection rate and be the infection rate in a (susceptible-infectious) contact, and let $K(d_{\mathrm{ij}}\,;\ )$ be the kernel function of the Euclidean distance between individual j and i (i.e., d_{ij}). The latent period (i.e., the sojourn time in class E) is assumed to be a two-parameter distribution $F_{\mathrm{u}}(\cdot)$ characterized by parameters and ^{2} (the density function is denoted by $f_{\mathrm{u}}(\cdot)$); similarly the infectious period (i.e., the sojourn time in class I) has density and distribution functions denoted as $f_{\mathrm{w}}(\cdot)$ and $F_{\mathrm{w}}(\cdot)$ respectively. Finally, let $\boldsymbol{\theta} = (\ ;\ ;\ ;\ ^2;\ ;\ ;\ )$ be vector of parameters in the model. As a result, we can express the likelihood function given the times of events as

$$\begin{aligned} L(\boldsymbol{\theta};\mathbf{E};\mathbf{I};\mathbf{R}) =& \{\prod_{\mathrm{j}\in\ _E^{-1}} \{\ +\sum_{\mathrm{i}\in \mathrm{I}_{E_j-}} K(d_{\mathrm{ij}}\,;\ )\} e^{-\mathrm{q}'_j}\} \times \prod_{\mathrm{j}\in\ _U} e^{-\mathrm{q}_{T_j}} \\ & \times \prod_{\mathrm{j}\in\ _I} f_{\mathrm{u}}(I_{\mathrm{j}} - E_{\mathrm{j}}\,;\ ;\ ^2) \times \prod_{\mathrm{j}\in\ _R} f_{\mathrm{w}}(R_{\mathrm{j}} - I_{\mathrm{j}}\,;\ ;\ ) \\ & \times \prod_{\mathrm{j}\in\ _{E\setminus I}} (1 - F_{\mathrm{u}}(t_{\max} - E_{\mathrm{j}}\,;\ ;\ ^2)) \times \prod_{\mathrm{j}\in\ _{I\setminus R}} (1 - F_{\mathrm{w}}(t_{\max} - I_{\mathrm{j}}\,;\ ;\ )) \end{aligned} \tag{5}$$

where $I(t)$ is number of individuals in class I at time t, and $_{\mathrm{E}}^{-1}$ is the set of individuals in class E excluding the index case, and $I_{\mathrm{t}-}$ is the set of individuals in I class (and not yet recovered) before time t. Also,

$$q'_{\mathrm{j}} = \int_{\mathrm{t}=0}^{\mathrm{e}_j} \{\ +\sum_{\mathrm{i}\in \mathrm{I}_{t-}} K(d_{\mathrm{ij}}\,;\ )\} dt; \tag{6}$$

and

$$q_{\mathrm{T}_j} = \int_{\mathrm{t}=0}^{\mathrm{t}_{max}} \{\ +\sum_{\mathrm{i}\in \mathrm{I}_{t-}} K(d_{\mathrm{ij}}\,;\ )\} dt: \tag{7}$$

2.1.1 Adaptation to Giant Hogweed Data

The likelihood function described above deals with a general S-E-I-R model which is used in our simulated example. In application to the giant hogweed data, we have instead fitted a S-I model. The likelihood function can be readily adapted by first omitting the terms corresponding to the sojourn time distributions. Two extra components are also modelled. First, the relative suitability of the sites is incorporated; second, as it is known that the survey was not extensive during the period when the snapshot was taken, we also estimate the probability that a colonized site reported at the third snapshot was actually non-reported at the second snapshot and denote it by p. Also, we remark that we have considered the likelihood for the observations at the last two snapshots (1987 and 2000) by conditioning on the occurrence of observations at the first snapshot (1970). The likelihood becomes

$$L(\boldsymbol{\theta};\mathbf{I}) = \{\prod_{\mathrm{j}\in\ _I^{-1}} b_{\mathrm{j}} \times \{\ c_{\mathrm{j}} + \sum_{\mathrm{i}\in \mathrm{I}_{\overline{j}-}} c_{\mathrm{j}}\, K(d_{\mathrm{ij}}\,;\ )\} e^{-\mathrm{q}'_j}\} \times \prod_{\mathrm{j}\in\ _U} e^{-\mathrm{q}_{T_j}} \tag{8}$$

where c_{j} is the suitability of the site j; $b_{\mathrm{j}} = 1$ if the colonized site is first reported at the second snapshot and $b_{\mathrm{j}} = p$ if the colonized site is first reported at the third snapshot.

2.2 Estimation

We estimate the parameters in a Bayesian framework by considering the joint posterior distribution of the model parameters given the data, which is given by

$$(\boldsymbol{\theta}|\mathbf{y}) \propto L(\boldsymbol{\theta};\mathbf{y})\ \ (\boldsymbol{\theta}); \tag{9}$$

where $(\boldsymbol{\theta})$ is joint prior distribution of the parameters and $L(\boldsymbol{\theta}; \mathbf{y})$ is the likelihood function with data $\mathbf{y}$ (observed and unobserved data). Markov Chain Monte Carlo (MCMC) methods are employed to estimate the joint posterior. In particular we use the (single-step) Metropolis-Hastings algorithm [3] and update the model parameters sequentially. We assume uniform priors for the parameters. To allow cryptic exposures in the simulation study, following Gibson & Renshaw [4], we adapt the reversible-jump algorithm [5] to the compartmental model setting. More details are given below.

2.2.1 Single-step MH algorithm for model parameters

I **Update , , , , ^{2}, , , and p sequentially**

(a) Propose a new parameter value, $'$, by performing a random-walk on the corresponding current value of the parameter, . Specifically,

$$' = + N(0; 1) \tag{10}$$

where $N(0; 1)$ is a random variate drawn from the standard normal distribution. If $' < 0$, it is rejected and the current value is retained. Note that a scaling constant can be multiplied with $N(0; 1)$ to control the step size and facilitate the convergence.

(b) Accept the proposed $'$ with probability

$$\frac{L(\boldsymbol{\theta}'; \mathbf{z})}{L(\boldsymbol{\theta}; \mathbf{z})} \tag{11}$$

where $'$ denotes the vector of parameters with replaced by $'$. Note that since we have used uniform priors and a symmetric distribution $N(0; 1)$ for the random walk, the acceptance probability reduces to the ratio of likelihoods.

(c) If $'$ is accepted, replace the current value by $'$, otherwise retains the current value.

(d) Apply the same algorithm to the remaining parameters sequentially.

II **Update the exposure times E_j**

(a) Randomly choose an exposure, j, and draw a new exposure time $E_j^{'}$ uniformly between $(0; t)$, where $t = I_j$ if j has become infectious, otherwise $t = t_{\max}$.

(b) With current data $\mathbf{z}$, accept the proposed new exposure time with probability

$$\frac{L(\boldsymbol{\theta}; \mathbf{z}')}{L(\boldsymbol{\theta}; \mathbf{z})} \tag{12}$$

where $\mathbf{z}'$ denotes the data with the current exposure time E_j replaced by $E_j^{'}$.

(c) If accepted, replace E_j with $E_j^{'}$, otherwise retain the current value.

2.2.2 Reversible jump algorithm for cryptic exposures

Sites that have been exposed but have not yet become infectious are referred to as cryptic exposures. In the simulation study in which a S-E-I-R model is fitted, we allow (unobserved) cryptic exposures and 'swap' of sites between the set $_{E\setminus I}$ and $_U$. These operations involve changes of model dimension, which requires the reversible jump algorithm. Adapted from Gibson & Renshaw [4], we apply two operations an *addition* and a *deletion* on the set of $_{E\setminus I}$. At each iteration during the MCMC run, each operation is equally likely to be applied.

I Addition of a cryptic exposure

(a) Randomly choose a site from $_U$ and move it to the set of $_{E\setminus I}$ and $_E$. Uniformly draw an exposure time $E_j^{'}$ between $(0; t_{\max})$ for this newly added cryptic exposure.

(b) Denote n_u and $n_{E\setminus I}$ as the number of sites in current sets $_U$ and $_{E\setminus I}$ respectively. Accept the proposed new sets and new exposure time with probability

$$\frac{L(\boldsymbol{\theta}; \mathbf{z}')}{L(\boldsymbol{\theta}; \mathbf{z})} \times \frac{n_u \times t_{\max}}{1 + n_{E\setminus I}} \tag{13}$$

where $\mathbf{z}'$ denotes the data with the changed sets ($_{E\setminus I}$ and $_E$ and $_U$) and with the current exposure time E_j replaced by $E_j^{'}$.

II Deletion of a cryptic exposure

(a) Randomly choose a site from $_{E\setminus I}$ (also from $_{E}$) and move it to the set of $_{U}$; delete the corresponding E_j accordingly.

(b) Accept the proposed new sets with probability

$$\frac{L(\theta;\mathbf{z}')}{L(\theta;\mathbf{z})} \times \frac{n_{E\setminus I}}{(1+n_u)\times t_{\max}} \tag{14}$$

where $\mathbf{z}'$ denotes the data with the changed sets ($_{E\setminus I}$ and $_{E}$ and $_{U}$) and with the current exposure time E_j being deleted.

2.3 Hypothesis Testing

For the hypothesis testing we use the *Anderson-Darling* test [6] which has the test statistics

$$A = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[(\ln Y_{(i)} + \ln(1-Y_{(n-i+1)})]; \tag{15}$$

where n is the sample size and $Y_{(i)}$ is the i^{th} largest sample. The null distribution of A [7] is used to compute the appropriate p-value using a package [8] available in the statistical software R. In view of our aim of detecting an anticipated mis-match in the tails of the distribution of imputed residuals, we do not adopt the commonly used K-S test which is known to be insensitive to the tail of the distribution.

2.4 Normalization of Spatial Kernel

As the secondary transmission rate and the kernel parameter are often highly correlated, normalization of the spatial kernel is implemented for reducing this correlation. Following [9, 10], we normalize the kernel by dividing $K(d_{ij}; \ )$ in the likelihood function by the normalization coefficient

$$S = \sum_{k\neq i;\, 1\le k\le N}^{N} K(d_{ik}; \ ): \tag{16}$$

Different normalization schemes may also relate to different assumptions over the dispersal mechanisms [9]. It is noted that in our setting above the spatial kernel describes how much more likely the colonizations of nearby sites are as compared with the colonizations of distant sites (in other words, there is competition among all possible infection links, which conforms to our model assumption), and this is referred to as *relative density dependence* [9].

3 Model Specifications for Simulated Example & Parameter Estimates

3.1 Model Specifications

Consider a population with size $N = 1000$ whose spatial coordinates are simulated uniformly on a square area 2000×2000. Assume all individuals in the population are susceptible at time 0 which is also the time of introduction of the force of primary infection , and assume that the epidemic is to be observed up to time $t_{\max} = 50$. For spatial kernels, we use an exponentially-bounded kernel (used to simulate the epidemic) $\exp(-\ _1 d)$ where d is the distance between the infectious and susceptible, and we have used a Cauchy-form kernel $1/(1+d/\ _2)$ and a power-law kernel d^{-k_3} as the incorrect kernels. The (correct) latent period distribution used to simulate the epidemic is Gamma distribution with mean and variance ^{2}; in fitting an incorrect latent period, an exponential distribution with mean is considered. A Weibull distribution, with *scale* parameter and *shape* parameters , is assumed for the infectious period. The locations and times of transitions from class S to class E are assumed to be unobserved, and the transitions from class E to I and the transitions from class I to R are assumed to be observed. However, we allow cryptic exposures (i.e., exposures that have not yet become infectious and remain undetected). Uniform priors, which should be constrained to bounded regions to ensure a proper posterior distribution, are specified for all parameters.

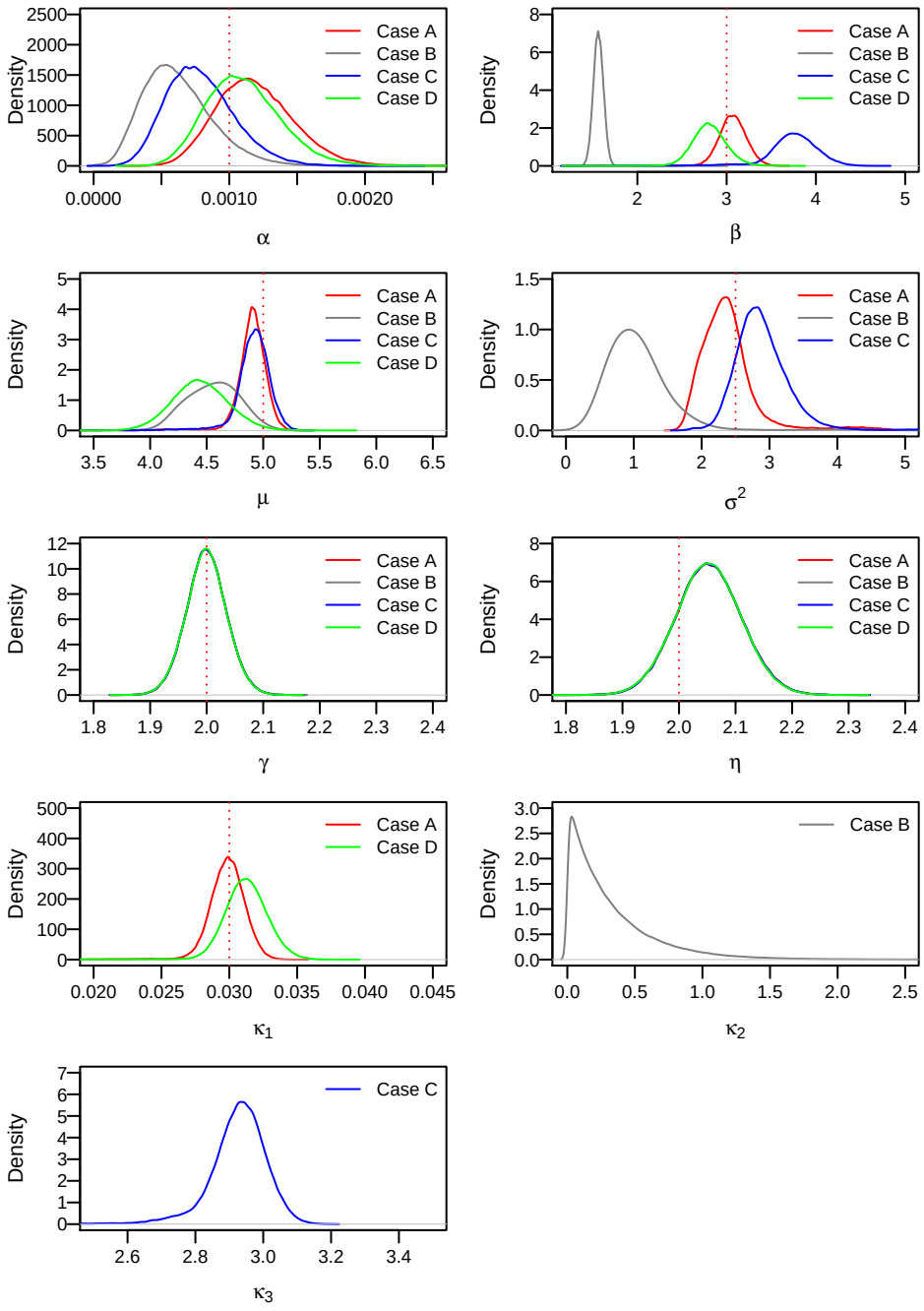

Figure S2: Posterior distributions of model parameters for models fitted to the simulated data (Replicate 1). Dotted lines are the actual values of the parameters used for simulating the epidemic.

3.2 Parameter Estimates

Figure S2 shows the posterior distributions of the model parameters for Case A to Case D from Replicate 1.

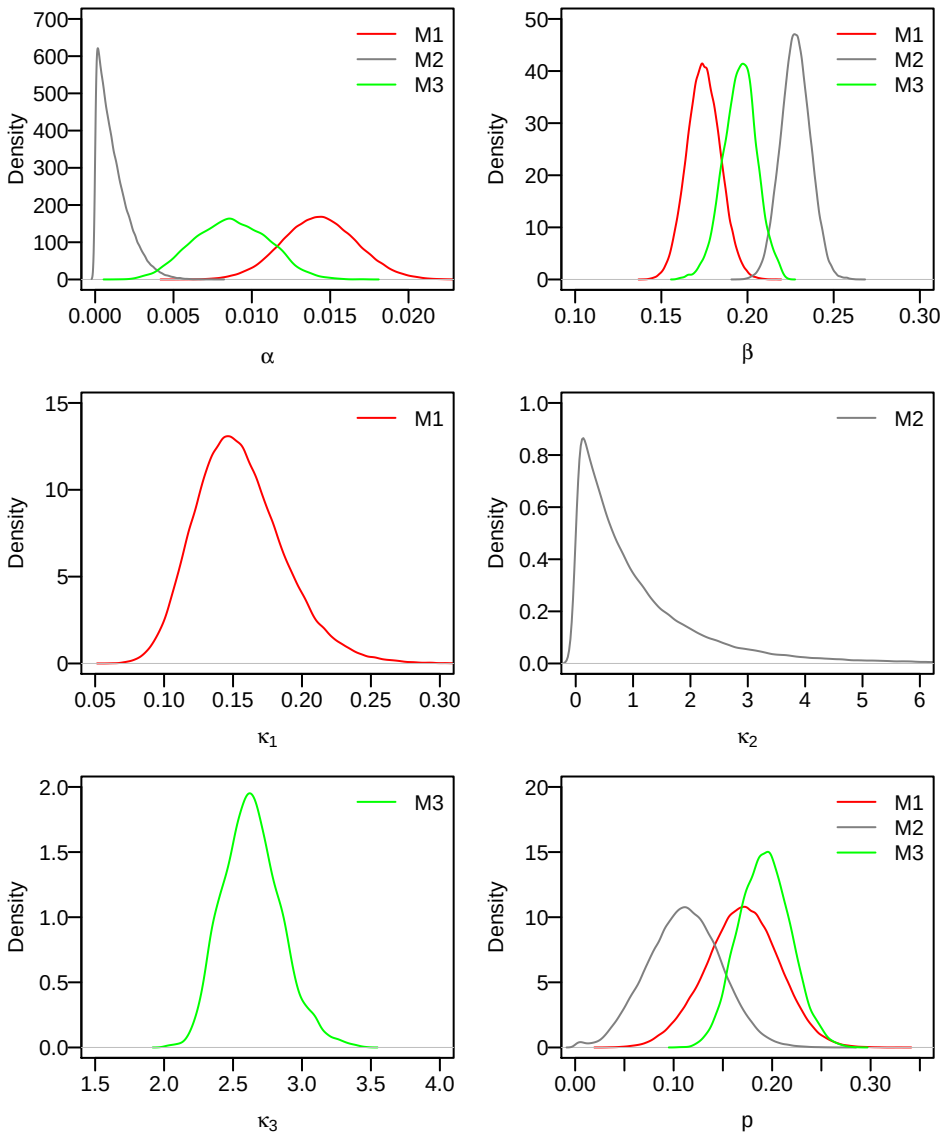

Figure S3: Posterior distributions of model parameters for models fitted to giant hogweed data where suitability of sites are considered.

4 Model Assumptions for Giant Hogweed Data & Parameter Estimates

4.1 Model Specifications

S-I models are fitted to the giant hogweed data. Similar to the simulated example, we fit an exponentially-bounded kernel $\exp(-\ _1d)$ (kernel A) and a Cauchy-form kernel $1=(1+d=\ _2)$ (kernel B) and a power-law kernel d^{-k_3} (kernel C) and compare their performance. The second snapshot at 1987 is known to be incomplete due to the insufficient efforts of surveying the sites. We therefore also estimate the probability, p, that a site has been colonized by 1987 but remained unreported given that it is reported at last snapshot (2000). Improper flat priors are used for all parameters.

4.2 Parameter Estimates

Figure S3 shows the posterior distributions of the model parameters from model M1 (kernel A) and model M2 (kernel B) and model M3 (kernel C) with the consideration of suitability.

5 Suitability of Sites

In the consideration of the suitability of sites, we adopt the mean relative suitability estimated from a previous study [10] in which an extensive range of covariates such as average temperature and altitude of sites was taken into account. To reinforce the confidence over the specification of the suitability, we subdivide the sites into three classes (assuming common suitability in each class) according to the estimated suitability from the analysis [10] and estimate the common suitability in each class.

Sites are classified into three classes (Less Favorable, Favorable and Highly Favorable, with suitability s_1, s_2 and s_3 respectively) according to the corresponding suitability estimated from the previous study [10]. We denote c_j as the estimate of suitability of site j given in [10]. If $0 < c_j \leq 0.25$, the site is classified as Less Favorable; if $0.25 < c_j \leq 0.5$, it is classified as Favorable; if $0.5 < c_j \leq 1.0$, it is classified as Highly Favorable. We estimate s_1 and s_2 as the suitability relative to $s_3 = 1$.

We consider fitting a model with kernel A (a 'better' kernel as we have shown) and we do not constrain the domain of parameters s_1 and s_2 (i.e., improper flat priors are used). Figure S4 shows the posterior distributions of parameters s_1 and s_2. From the figure, it is evident that $s_1 < s_2 < s_3 = 1$ and it indicates that the previous estimates of suitability [10] are reliable and broadly consistent with our estimates, which reinforces the confidence of adopting these earlier estimates [10] for our model.

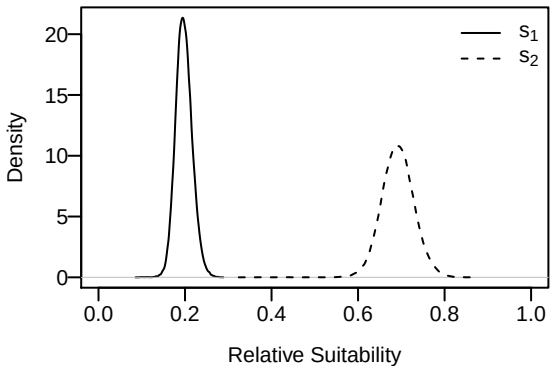

Figure S4: Posterior distributions of suitability parameters in the model (with kernel A) fitted to the giant hogweed data in which sites are classified into three classes

6 Posterior Predictive Checking Based on Spatial Autocorrelation Analysis

In this section, we consider posterior predictive checks based on spatial autocorrelation coefficients which measure spatial dependency among observations, we specifically consider two common measures Moran's I and Geary's c indexes [11]. The epidemic is simulated with kernel $K(d;\ ) = d^{-2.8}$. We respectively fit a correct kernel $K(d;\ _2) = d^{-k_2}$ (Model 1) and an incorrect kernel $K(d;\ _1) = \exp(-\ _1 d)$ (Model II) to the simulated data. Predictive distributions of the spatial autocorrelation coefficients from Model I and Model II, at three different time points within the observation period are compared to the corresponding measures computed from the actual (simulated) data.

We divide the 2000×2000 square area into $n = 100$ equally-sized square sub-regions and count the number of sites x_i in class I or class R in sub-region i at time points considered for the computation of Moran's I and Geary's c indexes

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{17}$$

$$c = \frac{n-1}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,(x_i - x_j)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{18}$$

where $\bar{x}$ is the mean of x_i over n sub-regions and w_{ij} is the spatial weight between sub-region i and j. There are many ways to define w_{ij}, and here we use the common binary weights in which $w_{ij} = 1$ if i is a *neighbour* of j,

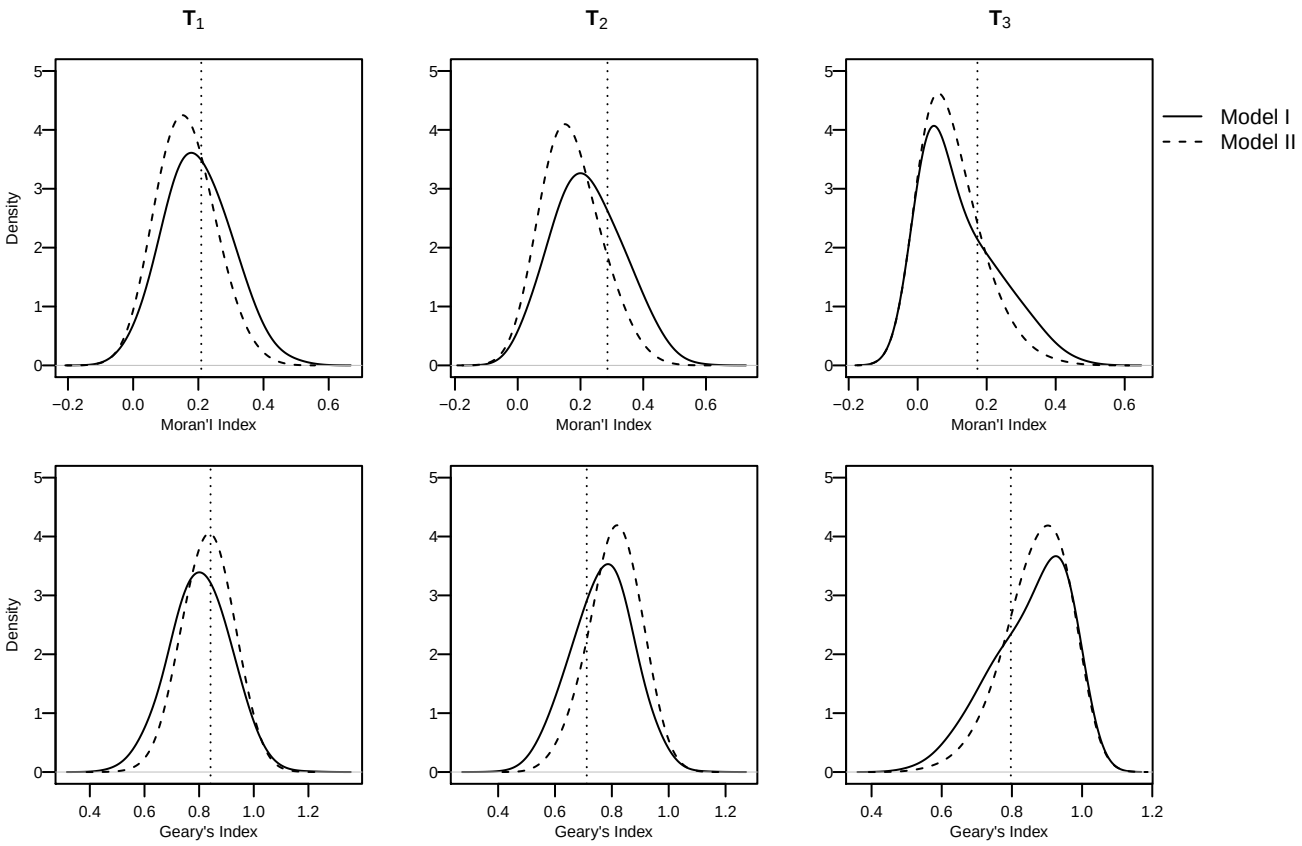

Figure S5: Posterior predictive distributions of Moran's I and Geary's c indexes obtained by simulating 1,000 epidemics from Model I and Model II respectively at time points $T_1 = 25$, $T_2 = 35$ and $T_3 = 45$. The vertical lines represent the observed values computed from the 'actual' epidemic.

otherwise $w_{ij} = 0$. The definition of *neighbour* is also open to variations, and here we define that any sub-regions whose centroids are within two sub-region width (400) from the centroid of sub-region i are considered to be the neighbour of i. Both indices are computed by using a package *spdep* [12] available in the statistical software R.

Model I represents a long-tail dispersal mechanism and Model II represents a localized dispersal mechanism - this is also illustrated by Figure S6 - and $(P(r_2) < 0.05|y)$ (i.e., the primary measure of degree of model mis-specification in utilizing r_2) shows strong evidence against Model II (90%) and no evidence against Model I. Figure S5 shows the predictive distributions of these two indices (at three different time points) obtained from simulating (1,000) epidemics respectively from Model I and Model II with the model parameters drawn from their respective posterior distributions. It can be seen from the figure that the posterior predictive distributions of the spatial autocorrelation indices from both models are broadly consistent with the observed values (i.e., all of the 95% two-sided intervals contain the actual value). This shows that posterior predictive checks based on these indices when only partially observed epidemic are available could be insensitive to the specification of the spatial kernel. Posterior predictive checking has to be computed 'offline' - for example, one needs to obtain the posterior distribution or point estimates of model parameters first and compute the required summary statistics based on simulation techniques. Our method, instead, can be easily embedded along with the estimation and by default takes the full posterior distribution of model parameters into account. It is also noted from above that summary statistics based on these spatial autocorrelation measures are subject to variations in definitions which might lead to different conclusions.

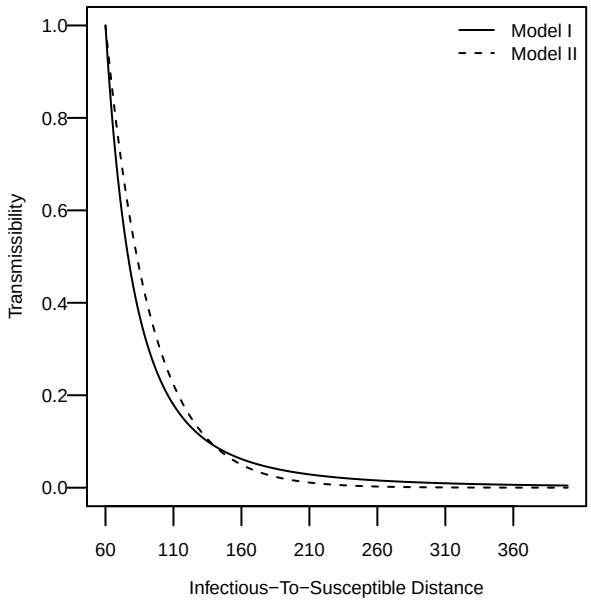

Figure S6: Estimated spatial kernels from fitting Model I and Model II with kernel parameters set to posterior means. Transmissibility is expressed relative to the amplitude of the respective kernel at $d = 60$ to highlight the difference between two kernels at both short and long distances.

7 Other Supplementary Figures

Figure S7 shows how the Great Britain is subdivided into 65 ring regions for the comparisons of colonized sites within the ring regions predicted by competing models (also see main text).

Figure S7: The partition of Great Britain according to intersection with 65 concentric annuli. Each annulus is centred on the black dot and has width 10km

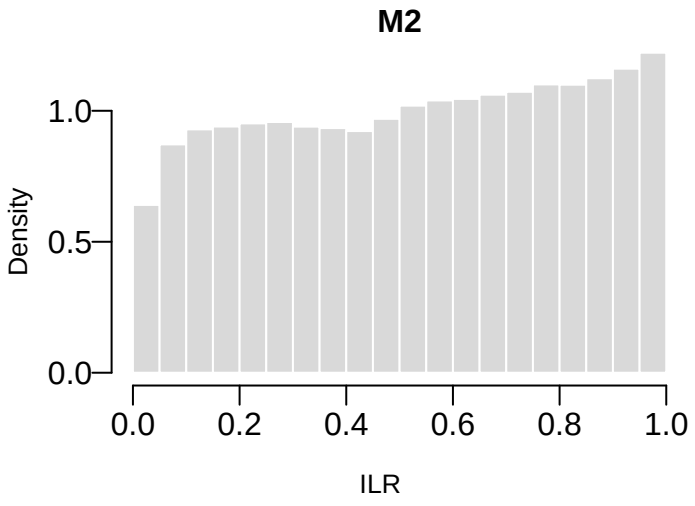

Figure S8: Distributions of subsets of imputed ILR which lead to p-values less than 0.05 from M2.

References

[1] Sellke T (1983) On the asymptotic distribution of the size of a stochastic epidemic. *Journal of Applied Probability* 390–394. (doi: 10.2307/3213811)

[2] Gibson GJ, Otten W, Filipe JA, Cook A, Marion G, Gilligan CA (2006) Bayesian estimation for percolation models of disease spread in plant populations. *Statistics and Computing* 16:391–402. (doi: 10.1007/s11222-006-0019-z)

[3] Chib S, Greenberg E (1995) Understanding the Metropolis-Hastings algorithm. *American Statistician* 49(4):327–335.

[4] Gibson GJ, Renshaw E (1998) Estimating parameters in stochastic compartmental models using Markov chain methods. *Mathematical Medicine and Biology* 15:19–40. (doi: 10.1093/imammb/15.1.19)

[5] Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82:711–732.

[6] Lewis PAW (1961) Distribution of the Anderson-Darling statistic. *The Annals of Mathematical Statistics* 1118–1124 (doi: 10.1214/aoms/1177704850)

[7] Marsaglia G, Marsaglia J (2004) Evaluating the Anderson-Darling distribution. *Journal of Statistical Software*, 9:1–5 *American Statistical Association*

[8] Bellosta CJG (2011) ADGofTest: Anderson-Darling GoF test, R package version 0.3, http://CRAN.R-project.org/package=ADGofTest

[9] Lindström T, Hákansson N, Wennergren U (2011) The shape of the spatial kernel and its implications for biological invasions in patchy environments. *Proceedings of the Royal Society B: Biological Sciences* 278(1711): 1564–1571. (doi: 10.1098/rspb.2010.1902)

[10] Catterall S, Cook AR, Marion G, Butler A, Hulme PE (2012) Accounting for uncertainty in colonisation times: a novel approach to modelling the spatio-temporal dynamics of alien invasions using distribution data. *Ecography* 35(10):901–911. (doi: 10.1111/j.1600-0587.2011.07190.x)

[11] Getis A (1991) Spatial interaction and spatial autocorrelation: a cross-product approach. *Environment and Planning A* 23(9):1269–1277.

[12] Bivand R (2013) spdep: Spatial dependence: weighting schemes, statistics and models, R package version 0.5-60, http://CRAN.R-project.org/package=spdep