Spatially informed back-calculation for spatio-temporal infectious disease models.

[1]*Gyanendra Pokharel and* [1]*Rob Deardon.*

[1] Department of Mathematics & Statistics, University of Guelph, Canada.

Abstract

Spatial dynamics are often key in understanding how infectious diseases spread. Since infection times are rarely observed in practice, however, identifying these dynamics can be difficult. Here, we implement a Bayesian hierarchical model in which spatial information is used to help infer the unobserved infection times, and through this, better identify the spatial-temporal transmission dynamics themselves.

Keywords: spatial disease transmission models; back-calculation; Bayesian hierarchical models; MCMC.

1. Introduction.

In many infectious disease systems, identifying spatial dynamics is of key importance. For example, in the UK 2001 foot mouth disease epidemic, distance between farms was shown to be a key risk factor for disease spread. Similarly, in many plant disease systems where spread occurs via wind-plume, rain or through the soil, distance is a driving risk factor. Spatio-temporal disease spread dynamics have been modeled by a number of authors (e.g. Deardon et al., 2010; Jewell et al., 2009). Unfortunately, identifying the space-time dynamics is made more difficult by the fact that infection times are rarely observed in practice, but must be inferred from disease reporting/diagnosis/observation times.

One way to deal with this is to explicitly model the incubation period; the time between infection and diagnosis. Typically, this is done in a Bayesian context, the (estimated) distribution of the incubation period being used to infer the infection times of individuals. This approach is known as back-calculation or back-propagation (e.g. Sweeting et al, 2005), and techniques such as data-augmented Markov chain Monte Carlo (MCMC) can be used to facilitate this approach.

However, standard back-calculation ignores the fact that in spatial systems, additional information on infection times may be derived from the proximity of individuals to each other. For example, if Euclidean distance is a key risk factor in the spread of a disease, then individuals close to each other are, on average, more likely to have been infected at points close in time, than individuals far away from each other.

To this end, we have extended the infectious disease models of Deardon et al., 2010 to incorporate a spatially informed infection time back-calculation model within a Bayesian hierarchical framework.

2. The Model.

2.1 Infectious Disease Model.

In the first instance, we assume a disease system in which distance is the key infection risk factor, and let P_{it}, the probability that susceptible individual i is infected at discrete time t, be given by:

$$P_{it} = 1 - \exp\left\{-\alpha\left(\sum_{j\in I(t)} d_{ij}^{-\beta}\right)\right\}, \quad \alpha, \beta > 0 \qquad [1]$$

where: α is an infectivity parameter that reflects the overall strength of the epidemic; β is a geometric spatial parameter; d_{ij} is the Euclidean distance between the individuals i and j; and I(t) is the set of infectious individuals at time t. Here, for simplicity, we will assume there is no latent period and that, once infected, individuals become immediately infectious for a known infectious period of γ times points before entering the 'removed state' (e.g. recovering with acquired immunity). This is known as an SIR framework.

Assuming known infection times for t = 1, 2,..., t_{max}, the likelihood is given by:

$$f(x|\alpha,\beta) = \prod_{t=1}^{t_{max}} \prod_{i\in S_{t+1}} (1 - P_{it}) \prod_{i\in I_{t+1}\setminus I_t} P_{it}$$

where: t_{max} is the last observed time point; x is the epidemic data; I(t+1) \ I(t) is the set of newly infected individuals at time point t+1; and S(t+1) is the set of disease susceptible individuals at time point t+1. The joint posterior is then given by combining the likelihood and the prior of the parameters $\theta = (\alpha, \beta)$ in the usual way.

2.2 Spatial Back-Calculation Model.

Let ψ_i, ϕ_i and λ_i be the infection time, diagnosis time and incubation periods for individual i, respectively. Then,

$$\phi_i = \psi_i + \lambda_i \qquad [2]$$

If we assume $\lambda_i = 0$ ($\phi_i = \psi_i$), then the model can be fitted in a Bayesian MCMC framework using the likelihood above with (likely) misspecified infection time data. However, noise introduced by the unaccounted for incubation period will tend to make the space-time dynamics harder to identify.

However, it is also possible to treat the infection times, ψ_i, to be random variables and augment the parameter vector θ with these variables. Standard back-calculation works by placing a prior on the incubation periods (λ_i); in this case, a discretized exponential distribution might be used:

$$\lambda_i \sim DExp(\zeta)$$

where ζ can be assumed known, or has a hyper prior placed upon it and is estimated. Infection times are thus inferred from the incubation periods and known observation times.

We can utilize spatial information in our data by letting $v_{ij} = |\psi_i - \psi_j|$, the difference in infection times between infectious individuals i and j, respectively, follow some prior distribution that is dependent on d_{ij}; for example, a Poisson prior with rate parameter δd_{ij} could be placed on the v_{ij}:

$$v_{ij} \sim Pois\ (\delta d_{ij}).$$

The purpose of this prior would be to force (with very small δ) or "encourage" (with larger δ) the infection events of individuals closer in space to be closer in time. Once again, δ can be assumed known, but here a hyper prior is placed upon it and it is treated as a parameter to be estimated.

3. Simulation Study.

A population of 100 individuals was constructed with the X and Y coordinates of individuals generated from a finite mixture of four equally weighted bivariate normal distributions in order to generate a 'realistic' clustered population. Specifically each individual's position was generated from:

$\binom{x}{y} \sim MVN\left(\binom{7.0}{7.0}, \begin{pmatrix} 3.5 & 0 \\ 0 & 4.5 \end{pmatrix}\right)$ with probability 0.25,

$\binom{x}{y} \sim MVN\left(\binom{4.0}{7.5}, \begin{pmatrix} 4.0 & 0 \\ 0 & 3.5 \end{pmatrix}\right)$ with probability 0.25,

$\binom{x}{y} \sim MVN\left(\binom{4.0}{2.5}, \begin{pmatrix} 3.5 & 0 \\ 0 & 4.5 \end{pmatrix}\right)$ with probability 0.25, or,

$\binom{x}{y} \sim MVN\left(\binom{7.5}{4.5}, \begin{pmatrix} 3.5 & 0 \\ 0 & 4.0 \end{pmatrix}\right)$ with probability 0.25.

Epidemics were simulated through this population using the spatial model of Equation (1) with parameters, $\alpha = 0.05$ and $\beta = 1.2$, with each individual having a fixed infectious period of $\gamma_I = 2$ time points. For each epidemic one individual was randomly infected at time t = 1, and then the simulation propagated until all individuals were infected, or the epidemic died out. Incubation periods (λ_i) were generated from an exponential distribution with mean 1, giving diagnosis times as shown in Equation (2). Ten epidemics were generated.

The standard model of Equation (1) was fitted to the diagnosis time data, and then the spatial back-calculation model of Section 2.2 was fitted, treating the infection times as unknown parameters using random-walk Metropolis Hastings MCMC.

Vague exponential priors were placed upon the two transmission model parameters, $\alpha \sim$ Exp (10^{-5}) and $\beta \sim$ Exp (10^{-5}). These priors insured positivity of the parameters, and were found to be imperceptibly informative.

For the spatial-back calculation model, δ was assumed to follow a positive half normal distribution; that is, $\delta \sim |N(0, \sigma_1^2)|$. Further, a hyperprior was placed on σ_1^2 such that $\sigma_1^2 \sim |N(0, 1.25)|$. Here, we assume we know the incubation period and set

$\lambda_i \sim$ Exp (1), although, of course, a hyperprior could be placed upon the incubation period distribution parameter.

4. Results.

Posterior mean estimates and 95% percentile intervals were obtained for each epidemic under each fitted model, and the results averaged over the ten epidemics are shown in Table 1.

Recalling that the true parameter values were $\alpha = 0.05$ and $\beta = 1.2$, we can see that under the standard model, wherein the misspecification of infection times is ignored, the average 95% posterior percentile intervals failed to capture the true values of the parameters. However, under the spatial back-calculation model, the true values were successfully captured by the intervals.

Table 1: Posterior estimates under standard and spatial back-calculation model.

Model	Parameter	Posterior Estimates	
		Mean	95% percentile intervals
Standard (2.1)	α	0.0354	(0.0266, 0.0465)
	β	0.8361	(0.2938, 1.0547)
Spatial Back-Calculation	α	0.0398	(0.0269, 0.0812)
	β	0.8399	(0.5663, 1.5912)
	δ	2.2049	(0.58156, 9.0581)
	σ_1^2	1.7375	(0.6772, 5.0878)

5. Discussion.

Due to space constraints, other results from a more in-depth simulation study, and application to experimental data from an experiment on tomato spotted wilt virus (TSWV), are not shown. However, the results shown are typical.

Further avenues for research might concern specifying the incubation period distribution and the spatial distribution placed on the ν_{ij}. An exponential distribution is typically used for incubation periods, but a more flexible distribution (e.g. gamma) could be considered. A myriad of possible spatial distributions could be considered. The Poisson distribution is limited since the variance and mean are equal, so a more flexible distribution such as a negative binomial could be considered. Further, a distribution that makes use of spatiotemporal information (i.e. information on the diagnosis times as well as location) could be used in the back-calculation. Additionally, our simplifying assumptions that we know the parameter of the incubation period distribution, the infectious period, and the fact that infectious period is the same for all infected individuals are generally unrealistic and will be relaxed in future work.

References

Deardon, R., Brooks, S., Grenfell, T., Keeling, M., Tildesley, M., Savill, N., Shaw, D., Woolhouse, M. (2010), "Inference for individual-level models of infectious diseases in large populations". In: Statistica Sinica, 20:239–261.

Jewell, C.P., Kypraios, T., Neal P., Roberts G.O. (2009), "Bayesian analysis for emerging infectious diseases". In: Bayesian Analysis, 4:191–222.

Sweeting, M., De Angelis, D., Aalen, O. (2005), "Bayesian back calculation using multi-state model with application to HIV." In: Statistics in Medicine, 24:3991–4007.