Spatio-temporal point processes, partial likelihood, foot and mouth disease

Peter J Diggle

Stat Methods Med Res 2006; 15; 325
DOI: 10.1191/0962280206sm454oa

Statistical Methods in Medical Research 2006; **15**: 325–336

Spatio-temporal point processes, partial likelihood, foot and mouth disease

Peter J Diggle Department of Mathematics and Statistics, Lancaster University, Lancaster, LAI 4YF, UK and Department of Biostatistics, Johns Hopkins University School of Public Health, USA

Spatio-temporal point process data arise in many fields of application. An intuitively natural way to specify a model for a spatio-temporal point process is through its conditional intensity at location x and time t, given the history of the process up to time t. Often, this results in an analytically intractable likelihood. Likelihood-based inference then relies on Monte Carlo methods which are computationally intensive and require careful tuning to each application. A partial likelihood alternative is proposed, which is computationally straightforward and can be applied routinely. The method is applied to data from the 2001 foot and mouth epidemic in the UK, using a previously published model for the spatio-temporal spread of the disease.

1 Introduction

In this paper, spatio-temporal point process data of the form (x_i, t_i): $i = 1, \ldots, n$, is considered, in which each (x_i, t_i) identifies the location and time of occurrence of an event of interest. It is assumed that the data consist of all relevant events in a pre-specified spatial region A and time interval $[0, T]$, that a parametric model for the underlying point process has been specified and that the goal is to make inferences about model parameters.

In Section 2, an approach to inference which uses a form of partial likelihood is developed.[1] The approach benefits from the generally attractive properties of likelihood-based methods while avoiding the computational complexity of full-likelihood inference. In Section 3 data from the UK 2001 foot and mouth epidemic are analysed using a model proposed by Keeling *et al*.[2] Section 4 discusses the potential for wider application of the methodology.

2 A partial likelihood

2.1 Definition

It is assumed that the point process is orderly, meaning roughly that coincident points cannot occur; for a rigorous discussion, see, for example, Daley and Vere-Jones[3] (Ch. 2).

Address for correspondence: Prof Peter J Diggle, Department of Mathematics and Statistics, Lancaster University, Bailrigg, Lancaster, LAI 4YF. Email: p.diggle@lancaster.ac.uk

  10.1191/0962280206sm454oa

$\mathcal{H}_t$ denotes the complete history of the process up to time t and $\lambda(x, t|\mathcal{H}_t)$ the conditional intensity for an event at location x and time t, given $\mathcal{H}_t$.

For data $(x_i, t_i) \in A \times [0, T]$: $i = 1, \ldots, n$, with $t_1 < t_2 < \cdots < t_n$, in principle, the log-likelihood function can be expressed as

$$L(\theta) = \sum_{i=1}^{n} \log \lambda(x_i, t_i|\mathcal{H}_{t_i}) - \int_0^T \int_A \lambda(x, t|\mathcal{H}_t)\, \mathrm{d}x\, \mathrm{d}t \tag{1}$$

See, for example, Daley and Vere-Jones[3] (Ch. 13). Two major obstacles to the routine use of Equation (1) are that the form of the conditional intensity may itself be intractable and that even when the conditional intensity is available direct evaluation of the integral term in Equation (1) may be impractical. Monte Carlo methods are becoming more widely available for problems of this kind.[4,5] However, in practice, these methods often need careful tuning to each application and the associated cost of developing and running reliable code can be an obstacle to their routine use.

As an alternative, computationally simpler approach to inference for models which are defined through their conditional intensity, a partial likelihood is proposed, which is obtained by conditioning on the locations x_i and times t_i and considering the resulting log-likelihood for the observed time-ordering of the events $1, \ldots, n$. To allow for right-censored event times, the risk set at time t_i is denoted by $\mathcal{R}_i$; in the absence of censoring, $\mathcal{R}_i = \{i, i + 1, \ldots, n\}$. Now let

$$P_i = \lambda \frac{(x_i, t_i|\mathcal{H}_{t_i})}{\sum_{j \in \mathcal{R}_i} \lambda(x_j, t_i|\mathcal{H}_{t_i})} \tag{2}$$

Then, the partial log-likelihood is

$$L_p(\theta) = \sum_{i=1}^{n} \log P_i \tag{3}$$

The partial likelihood defined by Equations (2) and (3) is a direct adaptation to the space–time setting of the seminal proposal in Cox[6] for proportional hazards modelling of survival data; essentially, the same idea has been suggested in Møller and Sorensen and Lawson and Leimich.[7,8] As discussed in Cox,[6] estimates obtained by maximizing the partial likelihood inherit the general asymptotic properties of maximum-likelihood estimators, although they may entail a loss of efficiency by comparison with full maximum-likelihood estimation. Also, some parameters of the original model may be unidentifiable from the partial likelihood. The loss of identifiability can be advantageous if the non-identified parameters are nuisance parameters as often applies, for example, to the baseline hazard function in the classic proportional hazards model for survival data.[1] Otherwise, and again as exemplified by the proportional hazards model for survival data, other methods of estimation are needed to recover the unidentified parameters.[9]

2.2 Identifiability and efficiency

When the conditional intensity function can be expressed as

$$\lambda(x, t|\mathcal{H}_t) = \lambda_0(t)g(x, t|\mathcal{H}_t) \tag{4}$$

for some function $\lambda_0(t)$, it follows immediately from Equation (2) that the partial likelihood provides no information about $\lambda_0(t)$. Specifically, if $g(\cdot)$ in Equation (4)is indexed by parameters θ, then the partial log-likelihood is

$$L_p(\theta) = \sum_i \log g(x_i, t_i|\mathcal{H}_{t_i}) - \sum_i \log \left\{ \sum_{j \geq i} g(x_j, t_i|\mathcal{H}_{t_i}) \right\} \tag{5}$$

Depending on the form of the function $g(\cdot)$ in Equation (4), its parameters may or may not be identifiable from the partial likelihood. For example, if $g(x, t|\mathcal{H}_t) = g(x)$, so that the point process is a Poisson process within independent spatial and temporal components, then the time-ordering of the events is uninformative and the partial likelihood method fails. In an informal sense, the time-ordering is expected to be highly informative, and the partial likelihood method, consequently, to have high efficiency, for parameters which relate to space–time interaction in the underlying process. Differentiation of Equation (5) gives the observed information for θ as

$$I_p(\theta) = \sum_i \left\{ g_i^{-1} \frac{\partial^2 g_i}{\partial \theta^2} - g_i^{-2} \left(\frac{\partial g_i}{\partial \theta} \right)^2 \right\} - \sum_i \left\{ G_i^{-1} \frac{\partial^2 G_i}{\partial \theta^2} - G_i^{-2} \left(\frac{\partial G_i}{\partial \theta} \right)^2 \right\} \tag{6}$$

where $g_i = g(x_i, t_i|\mathcal{H}_{t_i})$ and $G_i = \sum_{j \geq i} g(x_j, t_i|\mathcal{H}_{t_i})$. For comparison, the full log-likelihood associated with Equation (4) is

$$L(\theta) = \sum_i \log \lambda_0(t_i) + \sum_i \log g_i - \int \lambda_0(t) J(t)\, \mathrm{d}t \tag{7}$$

where

$$J(t) = \int g(x, t|\mathcal{H}_t)\, \mathrm{d}x \tag{8}$$

and the corresponding observed information for θ, treating $\lambda_0(t)$ as known, is

$$I(\theta) = \sum_i \left\{ g_i^{-1} \frac{\partial^2 g_i}{\partial \theta^2} - g_i^{-2} \left(\frac{\partial g_i}{\partial \theta} \right)^2 \right\} - \int \lambda_0(t) \frac{\partial^2 J}{\partial \theta^2}(t)\, \mathrm{d}t \tag{9}$$

3 Application: the 2001 foot and mouth epidemic

Foot and mouth disease (FMD) is a highly infectious viral disease of farm livestock. The virus can be spread directly between animals over short distances in contaminated airborne droplets and indirectly over longer distances, for example, via the movement of contaminated material. The UK experienced a major FMD epidemic in 2001, which resulted in the slaughter of more than six million animals. Its estimated cost to the UK economy was around £8 billion.[10]

3.1 Data

Data from the two counties most severely affected by the epidemic, Cumbria in the north-west of England and Devon in the south-west, can be analysed. Because the two counties are geographically separated, it shall be treated informally as two replicates of a natural experiment, thus allowing to compare parameter estimates and pool as appropriate. Figure 1 shows the spatial distribution of susceptible farms in Cumbria at the start of the epidemic. Figure 2 shows the evolving pattern of incident and prevalent cases as a discrete-time sequence.

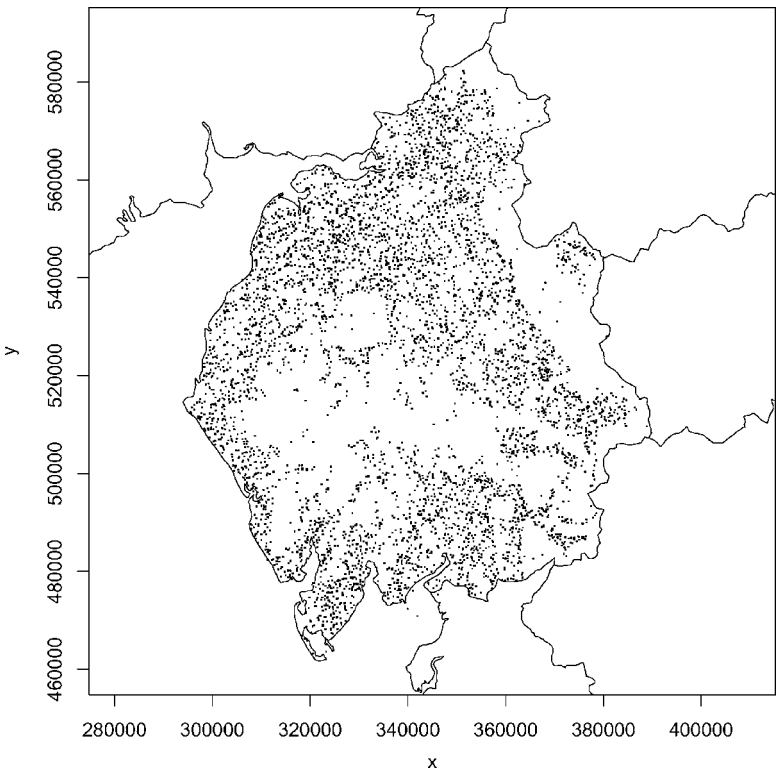

Figure 1 The spatial distribution of susceptible farms in Cumbria at the start of the 2001 foot and mouth epidemic.

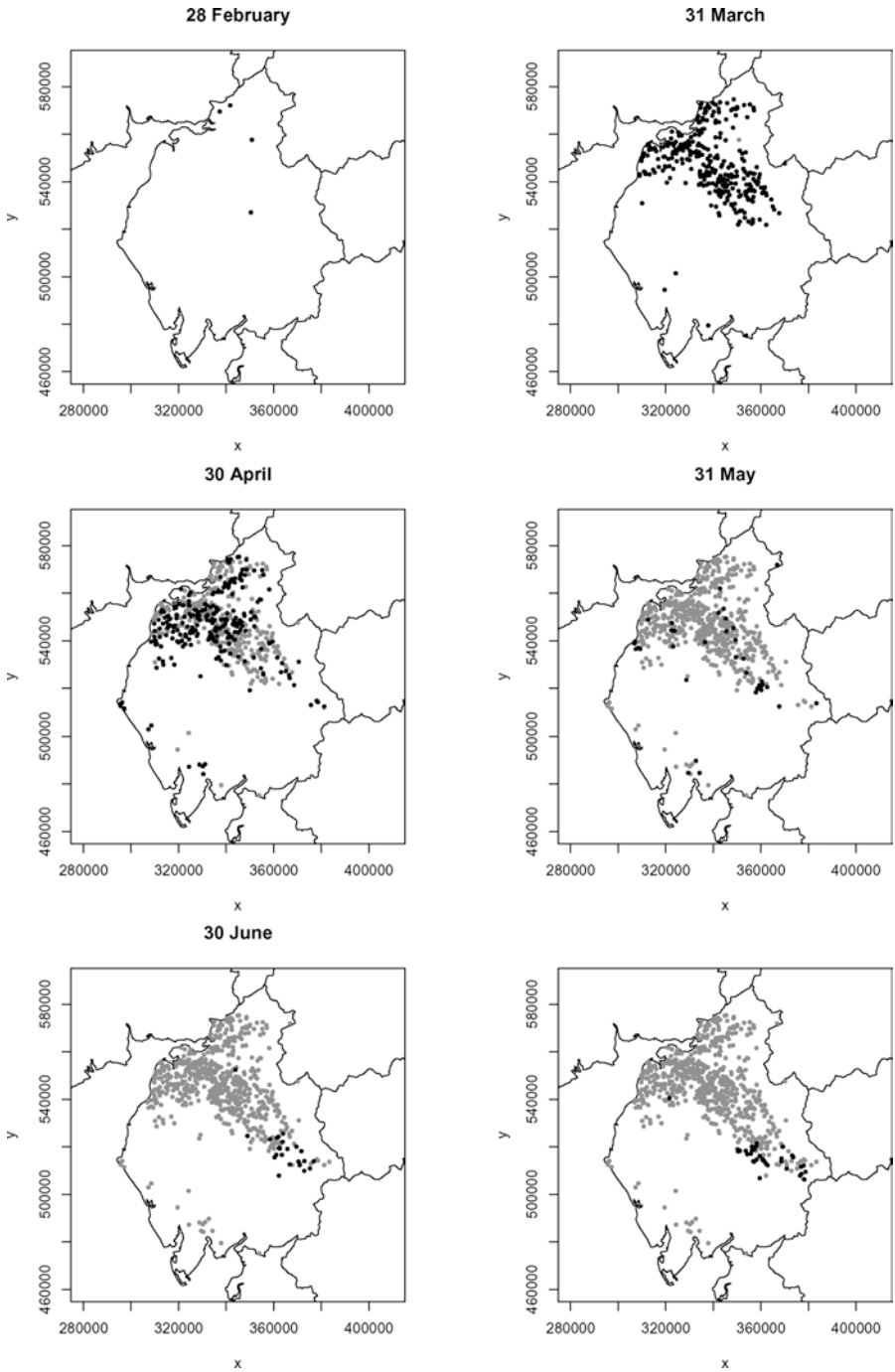

Figure 2 Incident and prevalent cases of the 2001 foot and mouth epidemic in Cumbria. Incident cases within the time interval between successive frames are shown in black. Prevalent cases on the dates indicated are shown in grey.

The spatio-temporal distribution of cases shows the typical pattern of an infectious disease process, with strong spatial aggregation of cases resulting from sequences of short-range transmissions between neighbouring farms, together with occasional, apparently spontaneous cases occurring at relatively long distances from all earlier cases. These effects become more obvious when the data are plotted dynamically as a spatio-temporal animation.

Information available on each farm includes the numbers of cattle and/or sheep held, the date, if any, on which the disease was reported and the date, if any, on which the farm's animals were slaughtered.

3.2 Model

The basic form of the model follows Keeling *et al*.[2] Section 3.4 discusses some possible extensions. Let $\lambda_{ji}(t)$ denote the conditional rate of transmission from farm j to farm i, given the history $\mathcal{H}_t$. Let n_{1i} and n_{2i} denote the numbers of cows and sheep held on farm i. Let $I_{ji}(t)$ denote an at-risk indicator for transmission of infection from farm j to farm i at time t; it is assumed that $I_{ji}(t) = 1$ if farm i is not infected and not slaughtered by time t, and farm j is infected and not slaughtered by time t. In the basic form of the model, it is assumed that the reporting date is the infection date plus a constant time τ, corresponding to the latent period of the disease plus any reporting delay. The complication that farm animals may be slaughtered after they become infected but before the disease is diagnosed is ignored. See the appendix for further discussion.

The central feature of the model is a transmission kernel,

$$f(u) = \nu \exp\left\{-\left(\frac{u}{\phi}\right)^{\kappa}\right\} + \rho \tag{10}$$

in which the powered exponential term corresponds to direct transmission of the infection over short distances, whereas the parameter ρ allows for occasional, apparently spontaneous cases occurring far from all currently infectious farms. For identifiability, $\nu = 1$, so that ρ in Equation (10) measures the relative importance of long-range transmission in the spread of the disease.

With the above definitions in place, the model specifies that

$$\lambda_{ji}(t) = \lambda_0(t) A_j B_i f(\|x_j - x_i\|) I_{ji}(t) \tag{11}$$

where $\lambda_0(t)$ is an arbitrary baseline hazard,

$$A_i = (\alpha n_{1i} + n_{2i}) \tag{12}$$

and

$$B_i = (\beta n_{1i} + n_{2i}) \tag{13}$$

The parameters α and β represent the relative infectiousness and susceptibility, respectively, of cows to sheep.

3.3 Fitting the model

For any farm i, the relevant conditional intensity is $\lambda(x_i, t_i|\mathcal{H}_{t_i}) = \sum_j \lambda_{ji}(t_i)$, and the partial log-likelihood follows by substitution of these conditional intensities into Equations (2) and (3). To maximize the partial log-likelihood, the Nelder–Mead simplex algorithm[11] is used, as implemented in the `R` functon `optim()`, which provides a numerical estimate of the Hessian matrix.

3.4 Results

In the model for the transmission kernel, the parameters κ and ρ are poorly identified because the cases which appear to correspond to long-range transmission are few in number and can be explained empirically either by including a small, positive value of ρ or by adjusting the value of κ. Because ρ corresponds formally to what is known to be a real effect, namely the indirect spread of infection via the movement of farm equipment and staff, ρ is retained as a positive-valued parameter to be estimated, but $\kappa = 0.5$ is fixed to correspond to the observation in Keeling *et al.*[2] that the transmission kernel is more sharply peaked than exponential.

It was first investigated whether the data in Cumbria and Devon support the assumption of a common set of parameters in the two counties. The likelihood ratio test statistic for common versus separate parameters is 2.98 on four degrees of freedom, hence $p = 0.56$ and therefore the hypothesis of common parameter values is accepted. The common parameter estimates $(\hat{\alpha}, \hat{\beta}, \hat{\phi}, \hat{\rho}) = (4.92, 30.68, 0.39, 9.9 \times 10^{-5})$ are then obtained. For all practical purposes, $\hat{\rho} \approx 0$, although a likelihood ratio test formally rejects $\rho = 0$ because the likelihood is sensitive to the precise probabilities which the model assigns to rare events.

One question of specific interest is whether the infectivities and susceptibilities for individual farms, A_i and B_i, are linear or sublinear in the numbers of animals. To investigate this, Equations (12) and (13) are extended to $A_i = (\alpha n_{1i}^{\gamma} + n_{2i}^{\gamma})$ and $B_i = (\beta n_{1i}^{\gamma} + n_{2i}^{\gamma})$, respectively, where γ is an additional parameter to be estimated. Fitting this five-parameter model results in a large increase in the maximized log-likelihood, from -6196.3 to -5861.4.

Another possible extension of the model would be to include farm-level covariates by defining $A_i = (\alpha n_{1i}^{\gamma} + n_{2i}^{\gamma}) \exp(z_i'\delta)$, where z_i is a vector of covariates for farm i, with a similar expression for the susceptibilities B_i. The z_i might, for example, codify management practices or other measured characteristics of individual farms which could affect their propensity to transmit, or succumb to, the disease. By way of illustration, adding a log-linear effect of farm area to the model is considered. The likelihood ratio statistic for the covariate effect is 3.26 on one degree of freedom, corresponding to $p = 0.07$. However, this test can be expected to be rather weak, because the observed distribution of farm area is extremely skewed, and the few farms with large areas will therefore have high leverage.

Estimates for the five-parameter model are shown in Table 1, together with approximate 95% confidence limits deduced from the numerical estimate of the Hessian matrix. Optimization was conducted on the log-scale for all parameters, which is why the confidence limits are not symmetric about the point estimates. Estimated correlations

Table 1 Parameter estimation for the five-parameter model fitted to combined data from Cumbria and Devon

Parameter	Estimate	95% confidence interval	
α	1.42	1.13	1.78
β	36.17	0.19	692.92
ϕ	0.41	0.36	0.48
ρ	1.3×10^{-4}	8.5×10^{-5}	2.1×10^{-4}
γ	0.13	0.09	0.21

among the parameter estimates were all small, the largest being 0.25 between $\log \phi$ and $\log \rho$. The results in Table 1 point strongly to a sub-linear dependence of infectivity and susceptibility on the numbers of animals. Note, however, that under the weak form of dependence implied by the estimate of γ, the parameter β is estimated very imprecisely.

These results are qualitatively similar to those reported in Keeling *et al.*,[2] although they only considered the case $\gamma = 1$. They reported point estimates $\alpha = 1.61$ and $\beta = 15.2$. They did not specify a parametric model for the transmission kernel but their Figure 1B shows similar behaviour to our fitted model, with a sharper-than-exponential mode at zero and a long upper tail. Their Figure 1B shows a decay from 1 at $u = 0$ to a value of approximately 0.1 at $u = 1$ km, compared with $\hat{f}(1) = 0.21$ in this paper.

Finally, a simple adaptation of the Nelson–Aalen estimator[9] (Ch. 4) is used to obtain a non-parametric estimate of the cumulative base-line hazard,

$$\hat{\Lambda}_0(t) = \int_0^t \hat{\lambda}_0(u)\, \mathrm{d}u.$$

Equation (11) is rewritten as $\lambda_{ji}(t) = \lambda_0(t)\rho_{ji}(t)$ and $\rho(t) = \sum_i \sum_j \rho_{ji}(t)$ is defined. The Nelson–Aalen estimator is now given by

$$\hat{\Lambda}_0(t) = \int_0^t \hat{\rho}(u)^{-1} \mathrm{d}N(u) = \sum_{i:t_i \le t} \hat{\rho}(t_i)^{-1}$$

where $\hat{\rho}(t)$ is the parametric estimate of $\rho(t)$ implied by the fitted model. Figure 3 shows the Nelson–Aalen estimates obtained from the Cumbria and Devon data. The generally lower estimates for Devon are consistent with the lower overall prevalence of the disease (137 cases out of 8182 at-risk farms in Devon and 657 cases out of 5090 at-risk farms in Cumbria). Both estimates are approximately linear over the first two to three months, by which time the epidemic in Devon has almost run its course. The slope of the Cumbria estimate increases thereafter. This does not necessarily imply a failure of the culling strategies being applied, as the model already takes account of their effects, but rather suggests that external environmental effects, for example, the increase in animal movements outdoors in spring and summer, may have promoted an increase in the virulence of the disease process.

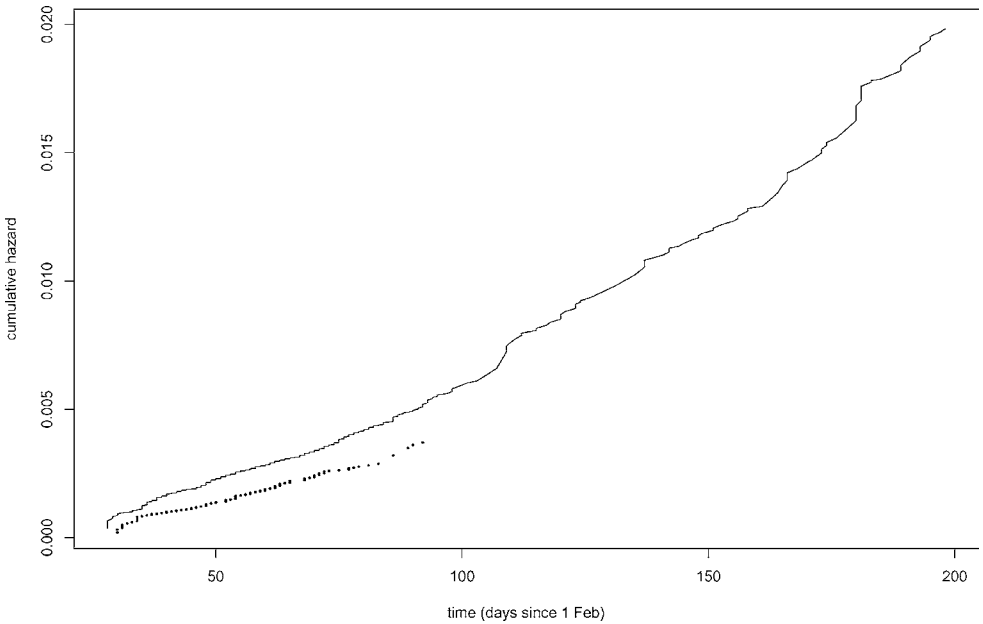

Figure 3 Estimated cumulative baseline hazards in Cumbria (solid line) and in Devon (dotted line) for the five-parameter model.

4 Discussion

The analysis of the FMD data has shown the feasibility of using the partial likelihood approach to answer a variety of questions relevant to an understanding of the disease process. A complete analysis of these data will be reported separately.

Statistical methods for the analysis of individual-level spatio-temporal data on the spread of infectious disease have become an active research area in its own right, as data-sets with the required level of detail become more widely available. For example, Neal and Roberts[12] fit a stochastic epidemic model to a classical data- set relating to a measles epidemic in an isolated village community.[13] They used Bayesian methods of inference implemented by means of a Markov chain Monte Carlo algorithm. Gibson,[14] motivated by applications in plant epidemiology, considered the problem of fitting continuous-time spatio-temporal stochastic models when the data consist of records of new infections in discrete time- intervals. Other examples include Lawson and Zhou,[15] who also analyse the Cumbria foot and mouth data, and Höhle *et al.*,[16] who consider both maximum likelihood and Bayesian methods of estimation for a stochastic model applied to data from a swine fever virus transmission experiment.[17]

From a methodological perspective, the approach is more widely applicable to spatio-temporal point process models and data for which the scientific focus is on the investigation of space–time interaction, whether this is modelled unconditionally as in a Poisson process with a non-separable space–time intensity function or conditionally as in models where the conditional intensity at time t depends explicitly on $\mathcal{H}_t$. To a

varying extent, maximum partial likelihood estimators may be less efficient than their maximum likelihood counterparts, but for many applications their implementation is much more straightforward. Ease of implementation is an important practical consideration when it is needed to explore the fit to a range of candidate models or to compare specific models within a given class.

Therefore, it is suggested that the partial likelihood approach has a potentially useful role to play in inference for spatio-temporal point process models which can be specified via their conditional intensity function. The approach is based on a generally accepted principle of inference with known asymptotic properties, although being computationally straightforward and therefore well-suited to routine use, provided, of course, that the parameters of interest remain identifiable. Work on the finite-sample properties of the method for a range of spatio-temporal point process models is in progress and will be reported in due course.

Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council through the award of a Senior Fellowship to Peter Diggle (Grant number GR/S48059/01) and by the UK Department for Environment, Food and Rural Affairs, grant number SE2930. I thank Laura Green, Matt Keeling, Graham Medley, Sir David Cox and Hugues Lassalle for very helpful discussions.

References

1 Cox DR. Partial likelihood. *Biometrika* 1975; **62**: 269–75.
2 Keeling MJ, Woolhouse MEJ, Shaw DJ, Matthews L, Chase-Topping M, Haydon DT, Cornell SJ, Kappey J, Wilesmith J, Grenfell BT. Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science* 2001; **294**: 813–7.
3 Daley DJ, Vere-Jones D. *An introduction to the theory pf point processes*. Springer, 1988.
4 Geyer C. Likelihood inference for spatial point processes. In Barndorff-Nielsen OE, Kendall WS, van Lieshout MNM eds. *Stochastic geometry: likelihood and computation*. Chapman and Hall, 1999: 79–140.
5 Møller J, Waagepetersen RP. *Statistical inference and simulation for spatial point processes*. Chapman and Hall, 2004.
6 Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B* 1972; **34**: 187–220.
7 Møller J, Sorensen M. Statistical analysis of a spatial birth-and-death process model with a view to modelling linear dune fields. *Scandinavian Journal of Statistics* 1994; **21**: 1–19.
8 Lawson AB, Leimich P. Approaches to space-time modelling of infectious disease behaviour. *Mathematical Medicine and Biology* 2000; **17**: 1–13.
9 Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical models based on counting processes*. Springer, 1992.
10 UK National Audit Office. *The 2001 outbreak of foot and mouth disease*. Report by the Comptroller and Auditor General, HC 939, Session 2001–2002. London: The Stationery Office, 2002.
11 Nelder JA, Mead R. A simplex algorithm for function minimisation. *Computer Journal* 1965; **7**: 308–13.
12 Neal PJ, Roberts GO. Statistical inference and model selection for the 1861 Hagelloch

measles epidemic. *Biostatistics* 2004; **5**: 249–61.
13 Pfeifsticker A. Beiträge zur Pathologie der Masern mit Besonderer Beruck- sichtgung der Statistischen Verhältmisse. MD Thesis, Eberhard-Karls Universitat, Tübingen, 1863.
14 Gibson GJ. Markov chain Monte Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. *Applied Statistics* 1997; **46**: 215–33.
15 Lawson AB, Zhou H. Spatial statistical modelling of disease outbreaks with particular reference to the UK FMD epidemic of 2001. *Preventive Veterinary Medicine* 2006, in press.
16 Höhle M, Jórgensen E, O'Neill PD. Inference in disease transmission experiments by using stochastic epidemic models. *Applied Statistics* 2005; **54**: 359–66.
17 Dewulf J, Laevens H, Koenen F, Vanderhallen H, Mintiens K, Deluyker H, de Kruif A. An experimental infection with classical swine fever in E2 sub-unit marker-vaccine vaccinated and in non-vaccinated pigs. *Vaccine* 2001; **19**: 475–82.

Appendix: Incubation, reporting delay and pre-emptive culling

How the partial likelihood analysis of the foot and mouth data would need to be modified to recognize both an incubation period and a reporting delay and to take account of pre-emptive culling is described now.

The current analysis assumes a fixed delay of τ days between the date of infection and the date of reporting. A more realistic assumption is to split this delay into an incubation period of τ_1 days between the date of infection and the date on which an infected farm becomes infectious, and a delay of τ_2 days between a farm becoming infectious and being reported as a case. Were it not for the existence of pre-emptive culling, the dates of infection and onset of infectiousness could be deduced from the reporting date and each farm could be treated as being infectious between τ_2 days before its reporting time and its culling time. Dealing with farms which are culled pre-emptively is more difficult because they may or may not have been infectious at the time of culling. The effect of this is that Equation (11) should be modified to

$$\lambda_{ji}(t) = \lambda_0(t) A_j B_i f(\|x_j - x_i\|) P_{ji}(t) \tag{14}$$

where $P_{ji}(t)$ is the probability that farm j is infectious and farm i susceptible at time t, given all of the available data. For any farm i, let T_i, S_i, R_i and C_i denote the times of infection, first infectiousness, reporting and culling, respectively, and as applicable. Formally, T_i can be thought as being arbitrarily large for a farm which never becomes infected and C_i as being arbitrarily large for a farm which is never culled.

Note that $S_i = T_i + \tau_1$ and $R_i = S_i + \tau_2$. To construct the partial likelihood, the conditional probabilities $P_{ji}(T_i)$ for all reported cases i need to be evaluated, and for all farms $j \neq i$ whether or not they are reported cases. For the partial likelihood, the notional time-ordering of non-case farms is irrelevant. The situations that need to be considered so as to deal with unreported cases are set out as shown below; clearly, all but one is straightforward.

$p_{ji}(T_i)$	$R_j < T_i + \tau_2$	$T_i + \tau_2 < R_j < \infty$	$R_j = \infty$
$C_j > T_i + \tau_2$	1	0	0
$T_i < C_j < T_i + \tau_2$	1	0	$p^*_{ji} = ?$
$C_j < T_i$	0	0	0

The quantity P^*_{ji} is the conditional probability, given the data, that farm i is infectious at time T_i but is never reported as a case because its notional reporting date R_j is greater than its culling date C_j. For this to happen, farm j would have to be infected within a time interval of length $\tau_2 - d_{ji}$, where $d_{ji} = C_j - T_i$; specifically, within the time interval from $T_i - \tau_1 - (\tau_2 - d_{ji})$ to $T_i - \tau_1$. On the assumption that censoring is uninformative, this gives $P^*_{ji} = 0$ for $d_{ji} < 0$ and $d_{ji} > \tau_2$, whereas for $0 < d_{ji} < \tau_2$,

$$P^*_{ji} = 1 - \exp\left(\int_{T_i-\tau_1-(\tau_2-d_{ji})}^{T_i-\tau_1} \lambda_j(u)\,\mathrm{d}u\right) \tag{15}$$

where $\lambda_j(u) = \sum_k \lambda_{kj}(u)$.

The derivation of Equation (15) assumes that the censoring, through pre-emptive culling, of unreported cases is non-informative. This would hold for a culling policy based only on farm location, for example, culling of all farms within a specified distance of a reported cases, but is dubious at best for other culling policies, such as culling of known dangerous contacts.

Even when censoring is non-informative, evaluation of Equation (15) is awkward because the term $\lambda_j(u)$ in the integrand which defines P^*_{ji} depends on quantities $\lambda_{kj}(u)$ as defined by Equation (14), which in turn involve the probabilities $P_{kj}(u)$.

A conservative strategy to deal with this difficulty would be to fit the model under two extreme scenarios, that is, for all pairs (i, j) such that $0 < d_{ji} < \tau_2$, set $p^*_{ji} = 0$ or $p^*_{ji} = 1$, respectively, and see whether the two scenarios give materially different fits.

A second possibility is to note that P^*_{ji} represents the probability of infection occurring in an interval of length $\tau_2 - d_{ji}$, given the prior constraint that it must occur within an interval of length τ_2 in order to have any chance of going unrecorded. Hence, a possible approximation would be to take

$$P^*_{ji} = 1 - \exp\left\{-c\frac{(\tau_2 - d_{ji})}{\tau_2}\right\}$$

for some positive constant c and to explore the sensitivity of the fit to different choices of c.

A third possibility is to construct a non-parametric estimate, $\hat{\lambda}_(t)$ say, of the time-varying unconditional rate of infection of a susceptible farm at time t and to replace the conditional intensity $\lambda_j(u)$ in Equation (15) by $\hat{\lambda}_(u)$ to give

$$P^*_{ji} = 1 - \exp\left(\int_{T_i-\tau_1-(\tau_2-d_{ji})}^{T_i-\tau_1} \hat{\lambda}(u)\,\mathrm{d}u\right)$$