ENVIRONMETRICS
*Environmetrics* 2005; **16**: 423–434
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/env.712

# Point process methodology for on-line spatio-temporal disease surveillance

Peter Diggle[1,2]*,†, Barry Rowlingson[1] and Ting-li Su[1]

[1]*Medical Statistics Unit, Lancaster University, Lancaster, UK*
[2]*Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA*

SUMMARY

We formulate the problem of on-line spatio-temporal disease surveillance in terms of predicting spatially and temporally localised excursions over a pre-specified threshold value for the spatially and temporally varying intensity of a point process in which each point represents an individual case of the disease in question. Our point process model is a non-stationary log-Gaussian Cox process in which the spatio-temporal intensity, $\lambda(x,t)$, has a multiplicative decomposition into two deterministic components, one describing purely spatial and the other purely temporal variation in the normal disease incidence pattern, and an unobserved stochastic component representing spatially and temporally localised departures from the normal pattern. We give methods for estimating the parameters of the model, and for making probabilistic predictions of the current intensity. We describe an application to on-line spatio-temporal surveillance of non-specific gastroenteric disease in the county of Hampshire, UK. The results are presented as maps of exceedance probabilities, $\mathrm{P}\{R(x,t)>c|\mathrm{data}\}$, where $R(x,t)$ is the current realisation of the unobserved stochastic component of $\lambda(x,t)$ and $c$ is a pre-specified threshold. These maps are updated automatically in response to each day's incident data using a web-based reporting system. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: Cox process; disease surveillance; gastroenteric disease; Monte Carlo inference; spatial epidemiology; spatio-temporal point process

## 1. INTRODUCTION

The AEGISS (Ascertainment and Enhancement of Gastrointestinal Infection Surveillance and Statistics) project aims to use spatio-temporal statistical methods to identify anomalies in the space–time distribution of non-specific, gastrointestinal infections in the UK, using the Southampton area in southern England as a test case. In this article, we use the AEGISS project to illustrate how spatio-temporal point process methodology can be used in the development of a rapid-response, spatial surveillance system.

Current surveillance of gastroenteric disease in the UK relies on general practitioners reporting cases of suspected food-poisoning through a statutory notification scheme, voluntary laboratory

*Correspondence to: P. Diggle, Medical Statistics Unit, Lancaster University, Lancaster, UK.
†E-mail: p.diggle@lancaster.ac.uk

Contract/grant sponsor: U.K. EPSRC; contract/grant number: GRS48059/01.
Contract/grant sponsor: U.S.A. National Institute of Environmental Health Science; contract/grant number: 1 R01 ES012054.
Contract/grant sponsor: Food Standards Agency, U.K.; NHS Executive Research and Knowledge Management Directorate.

*Received 15 February 2004*
*Accepted 15 September 2004*

reports of the isolation of gastrointestinal pathogens and standard reports of general outbreaks of infectious intestinal disease by public health and environmental health authorities. However, most statutory notifications are made only after a laboratory reports the isolation of a gastrointestinal pathogen. As a result, detection is delayed and the ability to react to an emerging outbreak is reduced. For more detailed discussion, see Diggle *et al*. (2003).

A new and potentially valuable source of data on the incidence of non-specific gastro-enteric infections in the UK is NHS Direct, a 24-hour phone-in clinical advice service. NHS Direct data are less likely than are reports by general practitioners to suffer from spatially and temporally localized inconsistencies in reporting rates. Also, reporting delays by patients are likely to be reduced, as no appointments are needed. Against this, NHS Direct data sacrifice specificity. Each call to NHS Direct is classified only according to the general pattern of reported symptoms (Cooper *et al*., 2003).

The current article focuses on the use of spatio-temporal statistical analysis for early detection of unexplained variation in the spatio-temporal incidence of non-specific gastroenteric symptoms, as reported to NHS Direct.

Section 2 describes our statistical formulation of this problem, the nature of the available data and our approach to predictive inference. Section 3 describes the stochastic model. Section 4 gives the results of fitting the model to NHS Direct data. Section 5 shows how the model is used for spatio-temporal prediction. The article concludes with a short discussion.

## 2. STATISTICAL FORMULATION

We define a *case* as any call to NHS Direct prompted by acute gastroenteric symptoms, indexed by date of call and residential location. The primary statistical objectives of the analysis are to estimate the 'normal' pattern of spatial and temporal variation in the incidence of cases, and to identify quickly any anomalous variations from this normal pattern. We address these objectives through a multiplicative decomposition of the space–time intensity of incident cases, with separate terms for: overall spatial variation, modelled non-parametrically as a smoothly varying surface $\lambda_0(x)$; temporal variation in the mean number of incident cases per day, $\mu_0(t)$, modelled parametrically through a combination of day-of-week and time-of-year effects; and residual space–time variation, modelled as a spatio-temporal stochastic process, $R(x,t)$. Hence, the spatio-temporal incidence is

$$\lambda(x,t) = \lambda_0(x)\mu_0(t)R(x,t)$$

Within this modelling framework, we define an *anomaly* as a spatially and temporally localised neighbourhood within which $R(x,t)$ exceeds an agreed threshold, $c$, and evaluate predictive probabilities $p(x,t;c) = \mathrm{P}\{R(x,t) > c|\text{data until time } t\}$. In practice, any anomalies identified by the analysis would become subject to follow-up investigations, including microbiologic analysis, in order to determine whether any form of public health intervention is warranted.

The analysis described in the present article uses NHS Direct data from the county of Hampshire, consisting of all 7126 cases reported between 1 January 2001 and 31 December 2002.

Because the pattern of calls to NHS Direct does not necessarily follow that of the overall population at risk, the use of census population counts to construct a baseline for local incidence could be misleading. We therefore use the accumulated historical pattern of incident cases to estimate the normal pattern of variation for the background spatial and temporal incidence rates. This relies on the fact that no major outbreak had been reported during the two-year period.

 

Our proposed model for space–time variation has a hierarchical structure, in the sense that it combines a model for a latent stochastic process, representing the unexplained space–time variation in incidence, with a model for the observed data conditional on this latent process. For Bayesian inference, we would add a third layer to the hierarchy, consisting of a prior distributional specification for the model parameters. In Bayesian terminology, the latent process is sometimes referred to as a parameter, and a model parameter as a 'hyper-parameter'. Whether or not we adopt the Bayesian viewpoint, an important difference between the two sets of unknowns is that model (or hyper) parameters are intended to describe *global* properties of the formulation, whereas the latent stochastic process describes *local* features.

In principle, we favour Bayesian predictive inference as a way of incorporating all sources of uncertainty into an assessment of predictive precision (see, for example, Diggle *et al*., 2003). However, in the current application, specifying the hyperprior for Bayesian inference is not very important given the correctness of the model. The reason is that our primary goal is predictive inference for the unobserved spatio-temporal process $R(x,t)$. Uncertainty in the predicted values of $R(x,t)$ reflects the sparseness of data on incident cases over the most recent few days, whereas estimation of global model parameters uses the relatively abundant data provided by the historical incidence pattern over a period of two years. It follows that prediction error will dominate estimation error, and predictive inference will therefore be relatively insensitive to the choice of prior. More pragmatically, a crucial requirement for the current application is that predictions can be updated daily. For daily updates of the predictive probabilities $p(x,t;c)$ we use a computationally intensive Markov chain Monte Carlo algorithm with parameters fixed at their estimated values, which runs overnight in our current computing environment.

## 3. MODEL FORMULATION

Our point process model is a straightforward adaptation of the model proposed by Brix and Diggle (2001), which in turn is an example of a spatio-temporal Cox process (Cox, 1955). Conditional on an unobserved stochastic process $R(x,t)$, cases form an inhomogeneous Poisson point process with intensity $\lambda(x,t)$, which we factorize as

$$\lambda(x,t) = \lambda_0(x)\mu_0(t)R(x,t) \tag{1}$$

In (1), $\lambda_0(x)$ represents purely spatial variation in the intensity of reported cases. Similarly, $\mu_0(t)$ represents temporal variation in the spatially averaged incidence rate. For identifiability, we scale $\lambda_0(x)$ to integrate to 1 over the study region, so that $\mu_0(t)$ describes the temporal variation in the mean number of incident cases per day. Each of these deterministic components of the model combines aspects of the underlying population at risk and of the pattern of disease. For example, if particular parts of the study region consistently report higher or lower incidence than the overall average, then this variation will be absorbed into $\lambda_0(x)$ and will not be identified as anomalous. Also, $\mu_0(t)$ includes both day-of-week effects, which to some extent are artefactual, and seasonal effects, which reflect genuine temporal variation in disease incidence. This emphasises that our surveillance system is designed to detect only spatially and temporally localized anomalies.

The remaining term $R(x,t)$ on the right-hand side of (1) is modelled as a stationary, unit-mean log-Gaussian stochastic process; hence

$$R(x,t) = \exp\{S(x,t)\} \tag{2}$$

 

where $S(x,t)$ is a stationary Gaussian process with mean $-0.5\sigma^2$, variance $\sigma^2$ and correlation function $\rho(u,v) = \mathrm{Corr}\{S(x,t), S(x-u,t-v)\}$. For a general discussion of log-Gaussian Cox processes, see Møller *et al*. (1998).

## 4. ESTIMATION

### *4.1. Overall spatial variation*

To estimate $\lambda_0(x)$, we use a kernel smoothing method with a Gaussian kernel, $\phi(x) = (2\pi)^{-1}\exp\{-0.5||x||^2\}$. The basic form of kernel estimation uses a fixed bandwidth $h > 0$ leading to the estimator

$$\tilde{\lambda}_0(x) = n^{-1}\sum_{i=1}^{n} h^{-2}\phi\{(x-x_i)/h\} \tag{3}$$

where $x_i, i = 1,\ldots,n$, are the locations of the $n$ incident cases in 2001 and 2002. Results using the kernel estimator (3) are reported in Diggle *et al*. (2003). We have since found that we obtain better results using an adaptive bandwidth kernel estimator, which takes the form

$$\hat{\lambda}_0(x) = n^{-1}\sum_{i=1}^{n} h_i^{-2}\phi\{(x-x_i)/h_i\} \tag{4}$$

The adaptive estimator (4) differs from (3) by allowing a different value of the bandwidth, $h_i$, to be associated with each observed case location $x_i$. This has the intuitively appealing consequence that it allows more smoothing to be applied to the data in sub-regions of relatively low intensity.

In our implementation we have used the adaptive bandwidth prescription

$$h_i = h_0\{\tilde{\lambda}_0(x_i)/\tilde{g}\}^{-0.5} \tag{5}$$

where $\tilde{\lambda}_0(x_i)$ is a pilot estimator of the form (3), $\tilde{g}$ is the geometric mean of the pilot estimates $\tilde{\lambda}_0(x_i)$ and $h_0$ is chosen subjectively (Silverman, 1986). In practice, we also apply an edge-correction as suggested by Diggle (1985) and Berman and Diggle (1989) to avoid substantial negative bias in $\hat{\lambda}_0(x)$ near the boundary of the study region.

We have compared the performance of the fixed and adaptive bandwidth versions of the kernel estimator on simulated realizations of inhomogeneous Poisson processes whose intensities are generated as $\lambda(x) = \exp\{S(x)\}$, where $S(x)$ is a stationary Gaussian process with covariance function $\mathrm{Cov}\{S(x), S(x-u)\} = \sigma^2\exp(-u/\phi)$. This model can generate a wide range of spatially aggregated point patterns by adjusting the values of the parameters $\phi$ and $\sigma^2$. The effect of increasing $\sigma^2$ is to generate higher peaks in the surface $\lambda(s)$, which leads to compact clusters of points. The effect of increasing $\phi$ is to make $S(x)$ more strongly spatially correlated, which leads to more slowly varying surfaces $\lambda(x)$ and correspondingly more diffuse aggregations of points.

We use the integrated squared error between the true and estimated intensity surfaces as a performance criterion. For each comparison, we simulate 100 samples, each consisting of 1000 points on a square region with side-length 89 units. From each simulated sample we compute the minimum integrated squared errors, $ISE_f$ and $ISE_a$, achieveable by the fixed and adaptive bandwidth kernel estimator, respectively, using the fact that the true $\lambda(x)$ is known for each simulated realization. We then compute $r = \log(ISE_a/ISE_f)$ as a measure of the comparative performance of the two estimators. To summarize the results for each pair

 

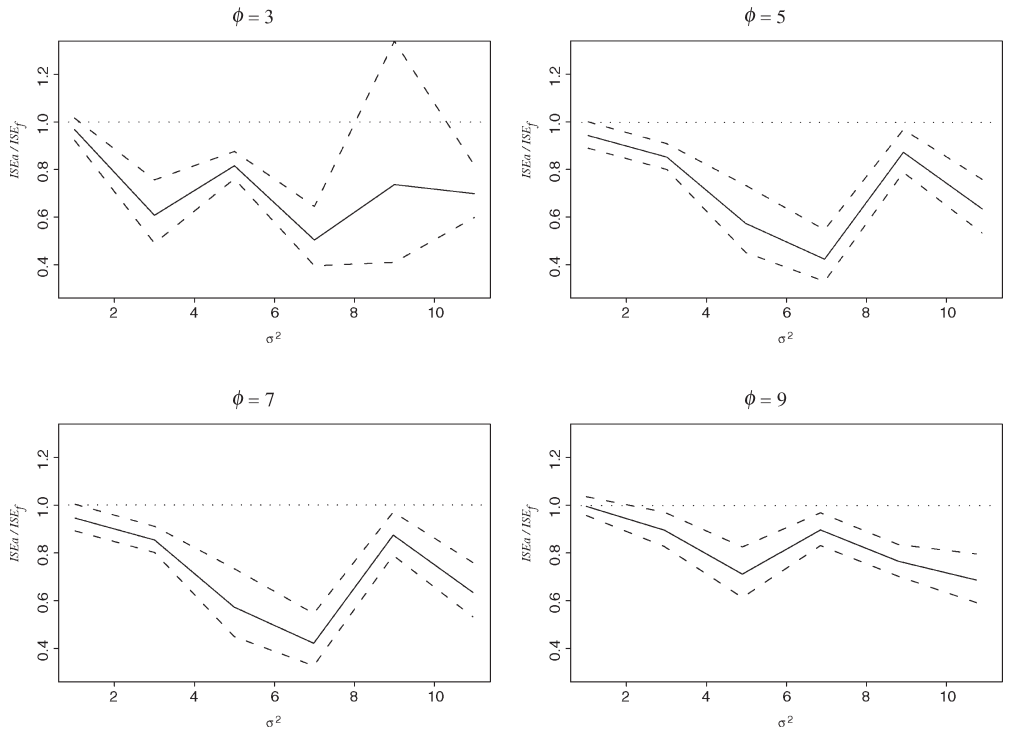


Figure 1. Summary results from simulation study to compare performance of adaptive and fixed bandwidth kernel estimators, for different values of the Gaussian process parameters $\sigma^2$ and $\phi$. The plotted lines show point estimates (solid line) and 95 per cent confidence limits (dashed lines) for the ratio of minimum integrated squared errors achievable by adaptive and fixed bandwidth estimators

of values of the model parameters $(\sigma^2, \phi)$, we compute means $\bar{r}$ and approximate 95 per cent confidence limits $\bar{r} \pm 2SE(\bar{r})$. Figure 1 shows the means and confidence limits back-transformed to the scale of ISE-ratios. These indicate the modest, but consistent, superiority of the adaptive over the fixed bandwidth kernel estimator. The superiority is more pronounced at medium to large values of $\sigma^2$, consistent with the fact that these are associated with more pronounced spatial heterogenity in the resulting point patterns. Within the range of our simulations, the effect of $\phi$ is less pronounced. We chose this range to span the varying degrees of spatial heterogeneity which we have experienced in our disease surveillance application. The adaptive kernel is favoured over the fixed kernel estimator under most, but not all, of the chosen scenarios. Note, in this context, that progressively increasing $\sigma^2$ will eventually produce a very light-tailed surface $\lambda(x)$ through the effect of the transformation from the Gaussian surface $S(x)$ to $\lambda(x) = \exp\{S(x)\}$. Under these conditions, there are theoretical reasons to believe that the adaptive kernel estimator will perform less well (Hall *et al.*, 1995).

Figure 2 shows our estimated surface $\hat{\lambda}_0(x)$ for the 2001 and 2002 NHS Direct data. This estimate uses the adaptive bandwidth prescription with $h_0 = 1.5$ km in (4), resulting in local values of $h_i$ ranging between 0.71 and 14.00.

### *4.2. Overall temporal variation*

With the scalings adopted for $\lambda_0(x)$ and for $R(x,t)$, the function $\mu_0(t)$ represents the unconditional expectation of the number of cases on day $t$. We therefore estimate $\mu_0(t)$ by a standard Poisson

 

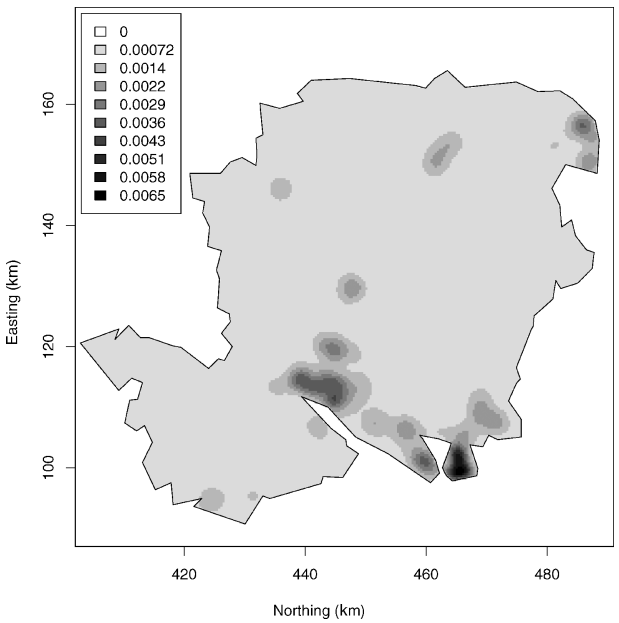


Figure 2. Kernel estimator for the overall spatial variation in reporting rates, $\hat{\lambda}_0(x)$, based on NHS Direct data from the county of Hampshire

log-linear regression model; note that the overdispersion induced by the stochastic component $R(x,t)$ does not affect the consistency of point estimates derived from the Poisson model, but does invalidate the nominal standard errors obtained under the Poisson assumption.

The empirical pattern of daily incident counts shows strong day-of-week effects, with excess numbers especially at weekends when more traditional sources of medical advice are less accessible. Time-of-year effects are also apparent, with higher incidence in the spring and autumn. Finally, there is an impression of an overall rising trend over time, which is likely to be due at least in part to progressive uptake of the NHS Direct service during its early years of operation. To take account of all of these effects, we fitted the model

$$\log\mu_0(t) = \delta_{d(t)} + \alpha_1\cos(\omega t) + \beta_1\sin(\omega t) + \alpha_2\cos(2\omega t) + \beta_2\sin(2\omega t) + \gamma t \qquad (6)$$

where $\omega = 2\pi/365$, corresponding to annual periodicity in incidence rates. Point estimates for the day-of-week effects in the regression model (6) are $\hat{\delta}_d = 2.24, 1.92, 1.76, 1.82, 1.76, 1.78, 2.12$, where $d = 1$ corresponds to Sunday, and so on. Point estimates of the harmonic regression parameters are $\hat{\alpha}_1 = -0.120$, $\hat{\beta}_1 = -0.083$, $\hat{\alpha}_2 = -0.013$ and $\hat{\beta}_2 = 0.054$, whilst the estimate of the slope parameter for the overall trend is $\hat{\gamma} = 0.00074$. Figure 3 compares the fitted regression curve with observed counts, averaged over successive one-week intervals to eliminate day-of-week effects.

### *4.3. Spatial and temporal dependence*

To estimate parameters of $S(x,t)$ we use the moment-based methods of Brix and Diggle (2001), which operate by matching empirical and theoretical descriptors of the spatial and temporal covariance structure of the point process model. For the current analysis, we assumed a separable correlation

 

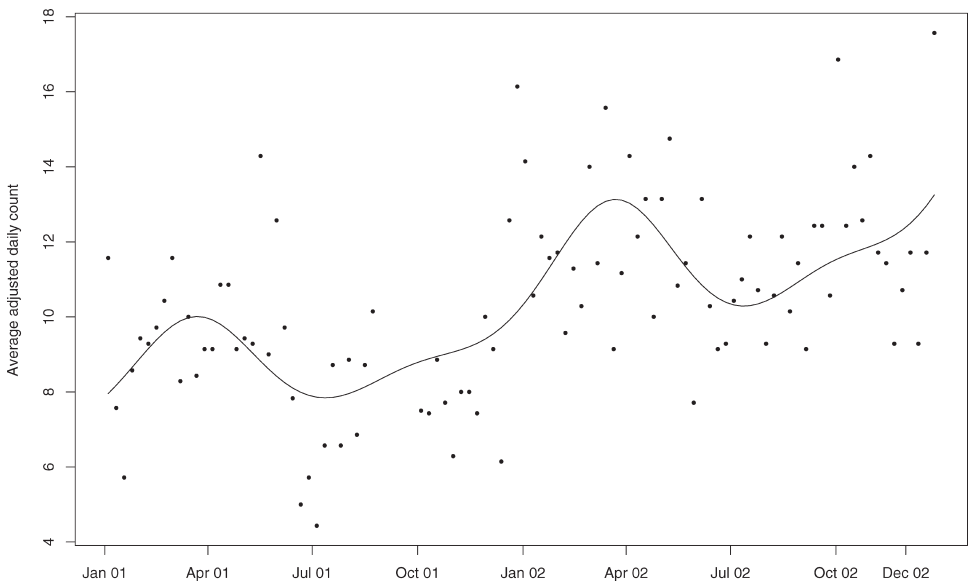


Figure 3. Observed counts of reported cases per day, averaged over successive weekly periods (solid dots), compared with the fitted harmonic regression model of daily incidence (solid line)

structure in which $\rho(u,v) = \rho_x(u)\rho_t(v)$. For the spatial component we used an exponential correlation function, $\rho_x(u) = \exp(-|u|/\phi)$. The corresponding spatial pair correlation function is $g(u) = \exp\{\sigma^2\exp(-|u|/\phi)\}$. We estimate $\sigma^2$ and $\phi$ to minimize the criterion

$$\int_0^{u_0} [\{\log\, \hat{g}(u)\} - \{\log\, g(u)\}]^2 du \tag{7}$$

where $u_0 = 2\,\text{km}$ and $\hat{g}(u)$ is a non-parametric estimate of the pair correlation function. We use a time-averaged kernel estimator with Epanecnikov kernel function and bandwidth $h = 0.2$; hence

$$\hat{g}(u) = \frac{1}{[2\pi u T|W|]} \sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{i \neq j} \frac{K_h(u - \parallel x_i - x_j \parallel) w(x_i, x_j)}{\lambda_t(x_i)\lambda_t(x_j)} \tag{8}$$

In (8), each of the summations over $i \neq j$ refers to pairs of events occurring on the same day, $t$, $T$ is the study period, $W$ the study area, $K_h(u) = 0.75h^{-1}(1 - u^2/h^2)$ when $-h \leq u \leq h$ and zero otherwise, $w(\cdot)$ is Ripley's (1977) edge correction, and $\lambda_t(x) = \lambda_0(x)\mu_0(t)$ is the unconditional spatio-temporal intensity. The validity of this estimator relies on our assumption that the spatio-temporal covariance structure is separable.

Figure 4(a) shows a good fit of the fitted parametric model for $\log g(u)$ to the non-parametric estimate $\log \hat{g}(u)$ defined by (8). The estimated parameter values are $\hat{\sigma}^2 = 8.85$ and $\hat{\phi} = 0.19\,\text{km}$.

For the temporal correlation structure of $S(x,t)$, we again assume an exponential form, $\rho_t(v) = \exp(-|v|/\theta)$, and estimate $\theta$ by matching empirical and theoretical temporal covariances of

 

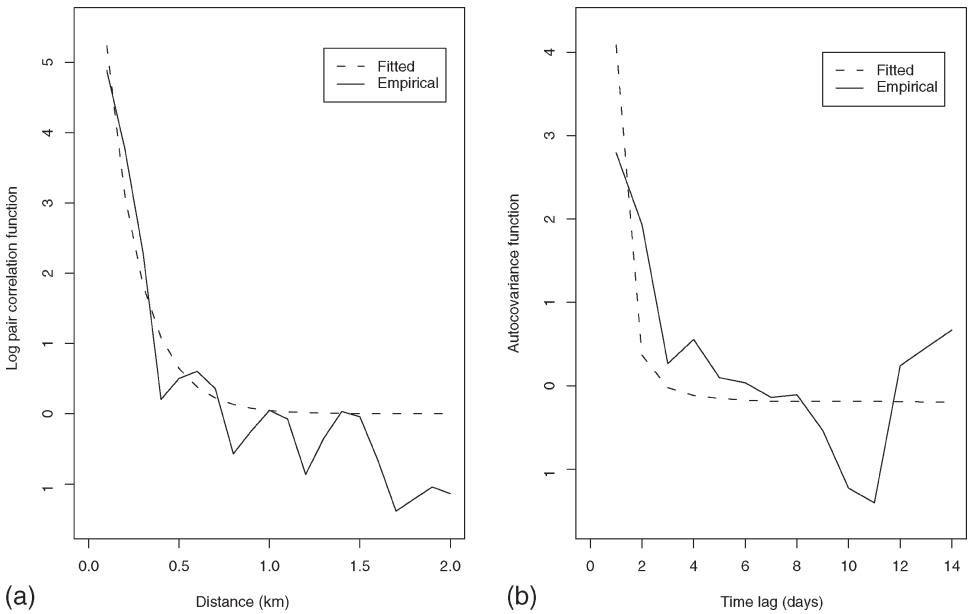


Figure 4. (a) Non-parametric (solid line) and fitted parametric (dashed line) log-pair correlation functions for the NHS Direct data. (b) Empirical (dashed line) and fitted (solid line) autocovariance functions for the NHS Direct data. See text for detailed explanation

the observed numbers of incident cases per day, $N_t$ say. For our model, the time variation in $\mu_0(t)$ makes the covariance structure of $N_t$ non-stationary. We obtain

$$\begin{aligned}\mathrm{Cov}(N_t, N_{t-v}) &= \mu_0(t)\mathbf{1}(v=0) + \{\mu_0(t)\mu_0(t-v)\} \\ &\quad \times \left\{ \int_W \int_W \lambda_0(x_1)\lambda_0(x_2)\exp[\sigma^2\exp(-v/\theta)\exp(-u/\phi)]\mathrm{d}x_1\mathrm{d}x_2 - 1 \right\} \qquad (9)\end{aligned}$$

Note that the expression for $\mathrm{Cov}(N_t, N_{t-v})$ given by Brix and Diggle (2001) is incorrect. Our estimation criterion for $\theta$ is

$$\sum_{v=1}^{v_0} \sum_{t=v+1}^{n} \{\hat{C}(t,v) - C(t,v;\theta)\}^2$$

where $v_0 = 14$ days and $\mathrm{C}(t,v;\theta) = \mathrm{Cov}(N_t, N_{t-v})$ as defined in (9). $\hat{C}(t,v)$ is the empirical autocovariance function, which is defined as

$$\hat{C}(t,v) = N_t N_{t-v} - \hat{\mu}_0(t)\hat{\mu}_0(t-v)$$

Figure 4(b) compares the empirical autocovariance function of the time series of daily incident cases $N_t$ with 'fitted' covariance functions obtained by averaging the values of $C(t,v;\hat{\theta})$ and $\hat{C}(t,v)$ over time, $t$, for each time lag, $v$. The estimated value of the temporal correlation parameter is $\hat{\theta} = 2.0$ days.

 

## 5. SPATIO-TEMPORAL PREDICTION

To solve the prediction problem of interest, namely the identification of spatially and temporally localized occurrences of unusually high incidence, we first need to generate a sample from the predictive distribution of the surface $S(x,t)$, and hence $R(x,t)$, conditional on the observed spatio-temporal pattern of incident cases up to and including time $t$. In practice, we do this on a fine grid of locations $x_k, k = 1,\ldots,m$, to cover the study region. As noted earlier, we also ignore uncertainty in the estimated values of the model parameters, on the grounds that, in this application, estimation uncertainty is negligible by comparison with prediction uncertainty. Having generated our sample, for each grid-point $x_k$ and a declared intervention threshold $c$, we approximate the predictive probability, $p(x_k,t;c) = \mathrm{P}\{R(x_k,t) > c|\mathrm{data}\}$, by the observed proportion of sampled values $R(x_k,t)$ which exceed $c$. We then plot these approximate exceedance probabilities as a colour-coded map, in which the colour scale is chosen so as to highlight only sub-regions where $p(x,t;c)$ is close to 1.

Following Brix and Diggle (2001), we use a Metropolis-adjusted Langevin algorithm (MALA) to generate samples from the predictive distribution of the current surface $S(x,t)$. Specifically, if $S_t$ denotes the vector with elements $S(x_k,t), k = 1,\ldots,m$, and $\mathcal{N}_t$ denotes the locations and times of all reported cases up to and including time $t$, the MALA generates samples from the conditional distribution of $S_t$ given $\mathcal{N}_t$.

Although the process $S_t$ is Markov in time, $\mathcal{N}_t$ is not, and the predictive distribution of $S(x,t)$ strictly depends on the complete history of $\mathcal{N}_t$. In practice, events from the remote past have a vanishing influence on the predictive distribution of $S(x,t)$. To avoid storing infeasible amounts of historical data, Brix and Diggle (2001) applied a 5-day cut-off, determined experimentally as the point beyond which retention of historical data had essentially no effect on the predictive distribution. The appropriate choice of cut-off will be application-specific, depending on the abundance of the data and the pattern of temporal correlation. In principle, a straightforward modification of the algorithm can be used to generate samples from the predictive distribution of $S(x,t+u)$ for any lead-time $u$. However, because of the short-range nature of the estimated temporal correlation, in our application forward projections rapidly become uninformative as the lead-time increases.

In our implementation of the MALA algorithm, we adjusted the variance of the proposal distribution to achieve an acceptance rate of around 0.57, as recommended in Roberts and Rosenthal (1998), and used block-upating to sample from the predictive distribution of $S(x,t)$ on each day, $t$, and a 256 by 256 grid of locations, $x$. Each day's predictive probabilities were computed as empirical proportions from a segment of 10 000 consecutive iterations. For a detailed description of the MALA algorithm, see Møller and Waagepetersen (2004).

For prediction using the NHS Direct data we fixed all of the model parameters at their estimated values, with the exception of the temporal trend parameter $\gamma$ in (6). This parameter was included in the model to allow for progressive uptake in the use of the NHS Direct service. On the assumption that the overall level of use has now stabilized, we chose to extrapolate the linear trend at a constant level $\hat{\gamma}t_0$, where $t_0$ corresponds to 31 December 2002. However, and as discussed in Section 6 below, the accuracy of this and other parametric assumptions can and should be reviewed periodically as data accumulate over time.

An integral part of the AEGISS project is to develop a web-based reporting system in which analyses are updated whenever new incident data are obtained. Each day, a program running in Lancaster checks for the arrival of new data. Whenever five consecutive days of data are identifed, these data are then passed to another program, which runs the spatial prediction algorithm. Outputs from the prediction algorithm in the form of maps of the exceedance probability surfaces $p(x,t;c)$ for

 

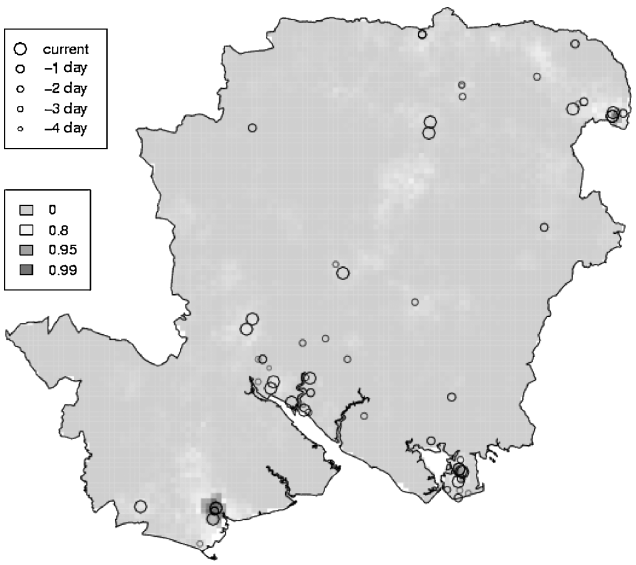


Figure 5. Posterior exceedance probabilities, $p(x,t;c) = \mathrm{P}[R(x,t) > c|\mathcal{N}_t]$, for $t$ corresponding to 6 March 2003 and $c = 4$

each of a set of values of $c$ are then passed back to a web-site. The actual analyses of the data are carried out using C programs with an interface to the R system (`http://www.r-project.org/`).

The threshold values used on the web-site are currently $c = 2$, 4 or 8. However, it would be preferable to relate these to the estimated parameters of the fitted model. Under our assumed model, the $p$-quantile of $R(x,t)$ is $c = \exp\{-0.5\sigma^2 + \sigma\Phi^{-1}(p)\}$. Setting $\sigma^2$ at its estimated value 8.85 would give threshold values $c = 0.54$, 1.60 and 12.13, corresponding to $p = 0.9$, 0.95 and 0.99, respectively.

Figure 5 shows a static example of the surface $p(x,t;c)$ for $t$ corresponding to 6 March 2003, and threshold value $c = 4$. The map suggests three possible anomalies near the south-west, south-east and north-east boundaries of the study region. In practice, it is more useful to track the evolution of $p(x,t;c)$ over successive days. An anomaly which appears one day and disappears the next is likely to be dismissed by a public health practitioner as a false positive, whereas one which persists over a few days, or at higher thresholds $c$, should prompt an intervention of some kind. The web-site `http://aegissdev.lancs.ac.uk:8080/Demo/` contains a record of daily updates over a three-month period, which can be examined interactively. Simple click operations allow the user to step forward and backward in time, and through the available values of $c$. These are currently set as $c = 2$, 4 and 8.

## 6. DISCUSSION

Point process modelling has the advantage that it does not impose artificially discrete spatial or temporal units on the underlying risk surface. Specifically, the scales of stochastic dependence in space and in time are determined by the data, and these estimated scales are then reflected in the amounts of spatial and temporal smoothing that are applied in constructing the predicted risk surfaces.

A possible objection to our particular model is that the Cox process is not a model for infectious disease. However, because of the duality between spatial clustering and spatial heterogeneity of risk noted by Bartlett (1964), our inhomogeneous Cox process model can describe clustered patterns of incidence empirically by ascribing local spatio-temporal concentrations of cases to peaks in the

 

stochastic process $R(x,t)$, after adjusting for overall spatial and temporal trends through the deterministic functions $\lambda_0(x)$ and $\mu_0(t)$. It is partly for this reason that we suggest using the term 'anomaly' rather than 'outbreak' to describe our findings, as we recognize that some anomalies will prove to be artefactual. In other words, we aim only to provide early indications of possible outbreaks, rather than definitive evidence that an outbreak has occurred.

Another possible concern is that our approach necessarily assumes that the residential location of each case is substantively relevant. In practice an individual's exposure to risk is determined by a complex combination of their residential, working and leisure locations and activities.

In specifying the spatial–temporal covariance function for $S(s,t)$, it is computationally advantageous to use a separable structure and the exponential form for the temporal correlation component, which makes the process Markovian in time. There is no corresponding advantage to using any particular form for the spatial correlation component, although coincidentally the exponential again provides a reasonable fit to the data.

Our current methods of parameter estimation, especially with regard to the spatial and temporal covariance parameters, are very *ad hoc*. We are adapting the methods described in Benes *et al*. (2002) and Møller and Waagepetersen (2004) to obtain maximum likelihood estimators of our model parameters. We are also conducting a simulation study of the area-wide sensitivity and specificity of the prediction algorithm by superimposing on the real data synthetic outbreaks of varying size and spatio-temporal extent.

The work reported here used data on cases reported up to the end of 2002. Examination of data subsequently obtained for 2003 illustrates the need for periodic review of the fitted model parameters. For example, Figure 6 shows an extrapolation of Figure 3, in which the model for the mean daily incidence, $\hat{\mu}_0(t)$, fitted from 2001 and 2002 data, has been projected forward in time and compared with

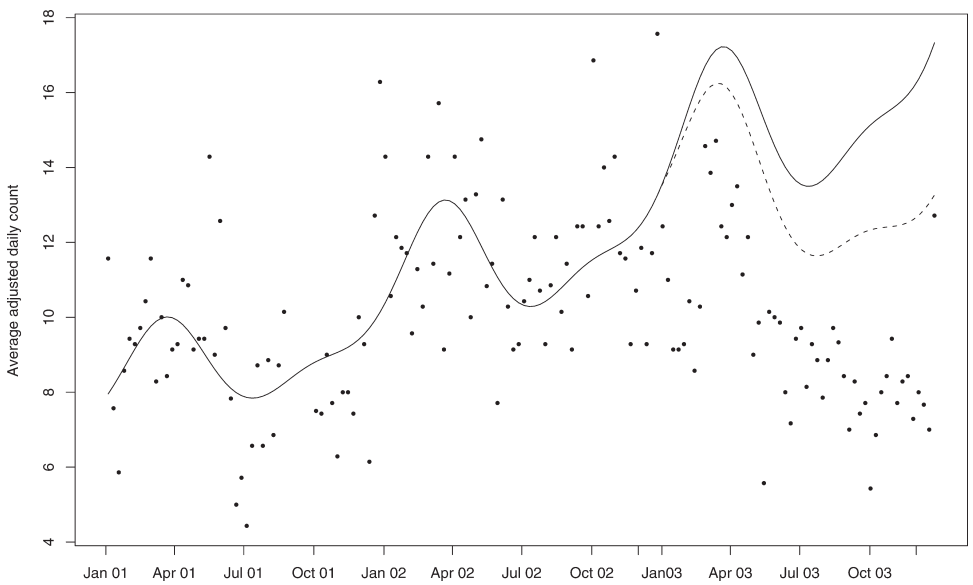


Figure 6. Observed counts of reported cases per day, averaged over successive weekly periods, for the years 2001 to 2003 (solid dots) compared with the harmonic regression model of daily incidence fitted to data from 2001 and 2002 only, extrapolated through 2003 using a continuation of the fitted linear term (solid line) and with the linear term extrapolated at constant level corresponding to 31 December 2002 (dashed line)

 

the actual 2003 data. The two projections correspond to continuation of the linear increase through 2003 and extrapolation of the linear term at a constant level. The actual 2003 data show the anticipated spring peak in incidence, but thereafter the incidence declines sharply by comparison with either of the extrapolated curves. To address this, we are investigating the use of a stochastic term in place of the deterministic linear trend component $\gamma t$ on the right-hand side of (6), as in West and Harrison (1997).

In conclusion, we have illustrated how spatial statistical methods can help to develop on-line surveillance systems for common diseases. The spatial statistical analyses reported here are intended to supplement, rather than to replace, existing protocols. Their aim is to identify, as quickly as possible, statistical anomalies in the space–time pattern of incident cases, which would then be followed up by other means. In some cases, the anomalies will be transient features of no particular public health significance. In others, the statistical early warning should help to ensure timely intervention to minimize the public health consequences—for example, when follow-up of cases in an area with a significantly elevated risk reveals exposure to a common risk factor or infection with a common pathogen.

## 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 U.S.A. National Institute of Environmental Health Science through Grant number 1 R01 ES012054, Statistical Methods for Environmental Epidemiology.

Project AEGISS is supported by a grant from the Food Standards Agency, U.K., and from the National Health Service Executive Research and Knowledge Management Directorate. We also thank NHS Direct, Hampshire and Isle of Wight, participating general practices and their staff, and collaborating laboratories for their input to AEGISS.

## REFERENCES

Bartlett MS. 1964. The spectral analysis of two-dimensional point processes. *Biometrika* **51**: 299–311.

Benes V, Bodlak K, Møller J, Waagepetersen R. 2002. Bayesian analysis of log Gaussian processes for disease mapping. Research Report No 3, February 2002, Centre for Mathematical Physics and Statistics, University of Aarhus, Denmark.

Berman M, Diggle P. 1989. Estimating weighted integrals of the second-order intensity of a spatial point process. *Journal of the Royal Statistical Society B* **51**: 81–92.

Brix A, Diggle PJ. 2001. Spatiotemporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society, B* **63**: 823–841.

Cooper DL, Smith GE, O'Brien SJ, Hollyoak VA, Baker M. 2003. What can analysis of calls to NHS Direct tell us about the epidemiology of gastrointestinal infections in the community? *Journal of Infection* **46**: 101–105.

Cox DR. 1955. Some statistical methods related with series of events (with Discussion). *Journal of the Royal Statistical Society, Series B* **17**: 129–157.

Diggle PJ. 1985. A kernel method for smoothing point process data. *Applied Statistics* **34**: 138–147.

Diggle PJ, Ribeiro PJ, Christensen O. 2003. An introduction to model-based geostatistics. In *Spatial Statistics and Computational Methods: Lecture Notes in Statistics*, Vol. 173, Møller J (ed.). Springer: New York; 43–86.

Diggle PJ, Knorr-Held L, Rowlingson B, Su T, Hawtin P, Bryant T. 2003. On-line monitoring of public health surveillance data. In *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*, Brookmeyer R, Stroup DF (eds). Oxford University Press: Oxford; 233–266.

Hall P, Hu TC, Marron JS. 1995. Improved variable window kernel estimates of probability density. *The Annals of Statistics* **23**(1): 1–10.

Møller J, Waagepetersen R. 2004. *Statistical Inference and Simulation for Spatial Point Processes.* Chapman & Hall: London.

Møller J, Syversveen A, Waagepetersen R. 1998. Log Gaussian Cox processes. *Scandinavian Journal of Statistics* **25**: 451–482.

Ripley BD. 1977. Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society B* **39**: 172–212.

Roberts GO, Rosenthal JS. 1998. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society, Series B* **60**: 255–268.

Silverman BW. 1986. *Density Estimation for Statistics and Data Analysis.* Chapman & Hall: London.

West M, Harrison J. 1997. Bayesian forecasting and dynamic models. In *Spatial Statistics and Computational Methods: Second edition*. Springer: New York.

 