STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:2469–2482
Published online 11 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2378

Additive models for geo-referenced failure time data

B. Ganguli[1,*,†] and M. P. Wand[2]

[1] *Department of Statistics, University of Calcutta, India*
[2] *Department of Statistics, University of New South Wales, Australia*

SUMMARY

Asthma researchers have found some evidence that geographical variations in susceptibility to asthma could reflect the effect of community level factors such as exposure to violence. Our methodology was motivated by a study of age at onset of asthma among children of inner-city neighbourhoods in East Boston. Cox's proportional hazards model was not appropriate since there was not enough information about the nature of geographical variations so as to impose a parametric relationship. In addition, some of the known risk factors were believed to have non-linear log-hazard ratios. We extend the geoadditive models of Kamman and Wand to the case where the outcome measure is a possibly censored time to event. We reduce the problem to one of fitting a Poisson mixed model by using Poisson approximations in conjunction with a mixed model formulation of generalized additive modelling. Our method allows for low-rank additive modelling, provides likelihood-based estimation of all parameters including the amount of smoothing and can be implemented using standard software. We illustrate our method on the East Boston data. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: additive models; disease mapping; failure time data; geostatistics; mixed models; penalized splines

1. INTRODUCTION

Modelling geographical variation of an outcome can be of interest in many areas of application. For example, increasing use of geographical information systems (GIS) worldwide has lead to a proliferation of spatial data for understanding environmental processes such as global warming. Similarly, in the field of disease mapping, geographical analysis of a health outcome can be useful for prioritizing the distribution of preventive services. We extend existing methodology [1] for modelling geographical variation of continuous outcomes which

[*] Correspondence to: B. Ganguli, Department of Statistics, University of Calcutta, 35, Ballygunge Circular Road Kolkata 700 019, India.
[†] E-mail: bgstat@caluniv.ac.in

Contract/grant sponsor: U.S. National Institutes of Health; contract/grant number: R01-ES10844-01

Received 24 March 2004
Accepted 10 August 2005

simultaneously allow for flexible modelling of non-linear covariate effects to the case when the outcomes are censored.

Non-parametric modelling of such failure-time data presents many additional challenges due to the need to account for censoring and the statistical challenges imposed by the presence of an unspecified baseline hazard. Smoothing martingale residuals have been used [2–5] to explore the correct functional form for a covariate but this method fails when covariates are correlated and 'adjusted variable plots' developed to address this problem can only validate if the assumption of linearity is correct. Low-rank smoothing splines have also been used [6, 7] as a flexible means of modelling unknown covariate effects on time to recurrence of breast cancer and develops penalized partial likelihood methods for estimation and inference. Literature on geospatial modelling of such data is limited. Existing methodology uses spatial frailties and Laplace approximations [8], hierarchical Bayesian models [9] and discrete time models with MCMC [9]. Our approach is based on that of Therneau and Grambsch [4, 5] who approximate the problem by a Poisson regression with an estimated offset. Advantages of this approach are cited below.

As indicated by the title, this paper is concerned with mapping a geographically referenced response variable. Bivariate smoothing is traditionally used for estimation in such cases as it allows us to avoid unduly restrictive parametric assumptions about the nature of the surface. The most prominent geostatistical method is kriging. However, mapping of geostatistical data is essentially equivalent to the bivariate smoothing problem in non-parametric regression where thin plate splines are the primary tools. Moreover, either of these approaches can be viewed as special cases of mixed model fitting [10–13]. The common thread running through all of these approaches is smoothing the geostatistical data through a linear combination of radial basis functions. Classical kriging and thin plate splines use n basis functions; where n is the number of points. However, recent literature in spline smoothing [10, 11, 14–16] has argued for $K \ll n$ basis functions on the grounds that there is little degradation in the functional fit. Such 'low-rank' approximations have an obvious computational advantage for large sample sizes.

If we assume that the approximation to the baseline hazard [4, 5] to be valid, our approach shares the following advantages with that of Kammann and Wand [1] for modelling continuous outcomes:

(1) a seamless means of modelling geographical variations in time to event while adjusting for the effect of other covariates;
(2) well studied in terms of parameter estimation and asymptotic standard errors. In particular, it provides likelihood-based estimates of the smoothing parameter;
(3) low-rank, as defined by Hastie [15, 16]; meaning that the number of basis functions used to construct the function estimates does not grow with the sample size. This is vitally important for disease mapping applications;
(4) implementable using standard software such as the `glimmix` macro in `SAS`, the `glmmPQL()` function in `R` and more recently, the `SemiPar` library in `R`.

The layout of this paper is as follows: Section 2 describes the application which motivated the development of this methodology. Sections 3 and 4 develop a low-rank additive mixed model for failure time data, first for a general case where it is necessary to model the effect of unknown covariates in a flexible but additive manner and then for the case of incorporating geographical covariates such as latitude and longitude in which case it is necessary to consider the problem of flexible *bivariate* modelling. Section 5 describes how our model formulation makes it is possible to use theory for generalized linear models to test for significance of covariates and model selection. Finally Section 6 presents an application of our methodology to the motivating data set.

2. DESCRIPTION OF THE APPLICATION AND DATA

Our motivating example is provided by the Maternal-Child Lung Study, a prospective study conducted by the East Boston Neighbourhood Health Centre to investigate risk factors affecting the age at which children from inner-city neighbourhoods develop wheezing and asthma. The outcome measure of interest was age at first physician-diagnosed asthma. Censoring occurred due to study termination or due to drop-out. The data set contains observations on the outcome and known risk factors for asthma such as age, gender, smoke exposure for 635 children from East Boston collected during regularly scheduled well-baby clinic visits and telephone interviews. It also contains observations on household locations measured in degrees latitude and longitude. These were obtained by geocoding residential addresses using the geographical information system `ArcView`.

In addition to studying the effect of traditional risk factors such as passive smoke exposure and maternal history of asthma, investigators were also interested in studying any residual geographical variations in the age of first onset of asthma. The basis for this is provided by the 'Life-Stress Paradigm' which postulates that geographical variations in the outcome after adjusting for the effect of known risk factors could reflect variations in the degree of community level stress factors such as poverty, unemployment and exposure to crime. While this is a little investigated area of asthma research, rising trends in asthma mortality in the United States have been found to disproportionately affect children living in inner-city areas. Carr and colleagues [17] documented up to tenfold differences in asthma hospitalization and mortality rates among some inner-city neighbourhoods relative to national rates. Marder and others [18] described excessive hospitalization and death rates from asthma in residents of inner-city Chicago. These variations do not appear to be the effect of differential access to health care since a Canadian study done in a setting with universal access to health care also found similar disparities. This suggests the potential for as-yet unidentified environmental risk factors affecting time to disease onset.

Study participants were largely white (50 per cent) and Hispanic (44 per cent). There was a roughly equal proportion of boys (49 per cent) and girls (51 per cent). Ages ranged from infancy to 6 years. Preliminary univariate analysis using Cox models for each covariate showed that many of the covariates were not significant predictors of the outcome. Table I lists all covariates that either exhibited some association in such univariate analysis or were considered as possible confounders based on prior substantive knowledge. We shall confine our attention to this subset of predictors in subsequent additive analysis. Since one of the predictors of interest was the number of lower respiratory tract infections in the first two years of life, we consider the subset of children whose age exceeded 2 years of age at first physician diagnosed asthma. Ideally such truncation of possible event times should be accounted for in the model but since this would considerably complicate the methodology while detracting from the main problem of modelling geographical variation, we do not do

Table I. Covariates that had some association with age at first physician-diagnosed asthma to a preliminary analysis.

Abbreviation	Description
Infant covariates	
`gender`	Indicator for male
`lri2`	Number of lower respiratory tract infections (LRI) in the first two years of life
Maternal covariates	
`race`	Categorical variable coded as White/Hispanic/other
`educ`	Maternal education Categorical variable coded as less than high school/high school/college/post college
`mat.asthma`	Indicator of whether mother ever had asthma
`ciguse`	Average daily number of cigarettes smoked by mother
`cotinine`	Maternal cotinine

The abbreviated names are used in the analysis summaries in Section 6.

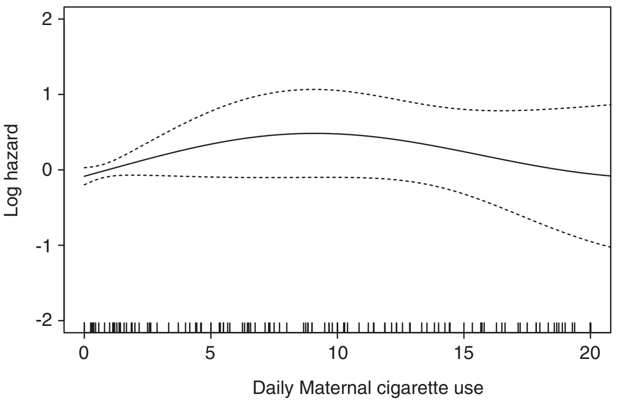

Figure 1. Estimated hazard ratio for average daily maternal cigarette use.

make any adjustments. It should be noted also that only 3 per cent of the children were aged less than 2 years at first occurrence of the event.

Subsequent univariate analysis to detect non-linearity in the effect of `ciguse`, using the methods of Therneau and Grambsch [5] and several choices of degrees of freedom selects a model with three degrees of freedom in `ciguse` as having the largest AIC. The corresponding plot is given in Figure 1.

Similar analysis fitting an additive model in latitude and longitude leads to a model with three degrees of freedom in each of latitude and longitude. However, an additive model in latitude and longitude is not a restriction that we would want to impose *a priori* and the next sections present a more general model formulation (Figure 2).

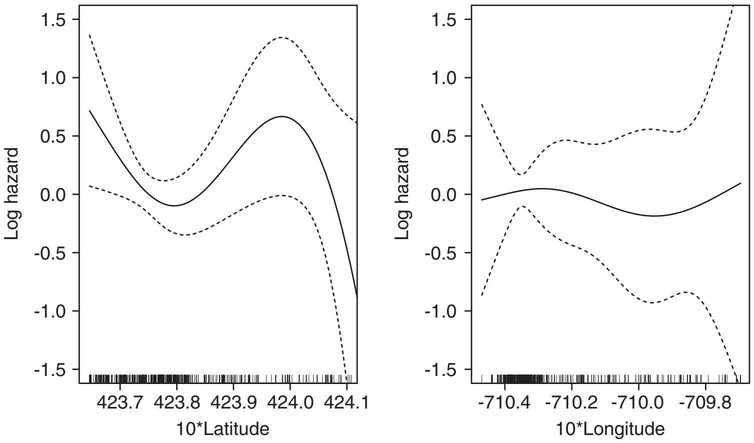

Figure 2. Estimated hazard ratio for latitude and longitude.

3. ADDITIVE MODELS FOR FAILURE TIME DATA

In this section, we shall first demonstrate the reduction of flexible additive modelling of covariate effects on failure time data to a generalized linear mixed model (GLMM). Our data consists of a random sample, (T_i, C_i) for $i = 1, 2, \ldots, n$. Here the T_i are non-negative continuous random variables denoting time to the outcome of interest measured from an appropriate starting point and the C_i are corresponding censoring times. We also have measurements on p covariates for each observation. For the sake of simplicity in notation, we consider the case $p = 2$ and denote the covariates by (V_i, W_i), $i = 1, 2, \ldots, n$. Extension of the formulation for a general value of p is straightforward. We assume a right-censoring scheme, i.e. the observed random sample is

$$(U_i, \delta_i, V_i, W_i)$$

where

$$U_i = \min(T_i, C_i) \quad \text{and} \quad \delta_i = I[T_i \leqslant C_i], \quad i = 1, \ldots, n$$

We also assume that censoring is non-informative, i.e. that T_i and C_i are conditionally independent given (V_i, W_i). This allows censoring to depend on covariates but, conditional on the covariate, censoring times do not provide any information about the distribution of failure times.

Finally, we shall assume that our data are generated from a distribution with hazard function of the form

$$\lambda(t|v, w) = \lambda_0(t) \exp\{f(v) + g(w)\}$$

where $\lambda(t|v, w)$, $t > 0$, denotes the hazard function at time t for an individual with covariate value $V = v$, $W = w$; $\lambda_0(t)$, $t > 0$ is the unspecified baseline hazard and $f(\cdot)$ and $g(\cdot)$ are unspecified but smooth functions. We assume here and in subsequent discussion that the log hazard function is an additive function of covariates. Our model assumes proportional hazards in the sense that all hazard ratios are time-invariant although we relax the assumption of linearity in the covariates [19].

Our approach is based on that of Therneau and Grambsch [4, 5] who note that, via martingale theory, we have the approximate Poisson regression model:

$$E(N_i|v_i,w_i)=\exp\{f(v_i)+g(w_i)\}\int_0^\infty Y_i(s)\lambda_0(s)\,\mathrm{d}s$$

where $N_i(t)=I[T_i\leqslant t,\delta_i=1]$ and $Y_i(t)=I[T_i>t]$ for $i=1,\ldots,n$. N_i refers to the counting process evaluated at $t=\infty$. The terms $\int_0^\infty Y_i(s)\lambda_0(s)\,\mathrm{d}s$ are unknown but following Grambsch, Therneau and Fleming [4], we shall plug in estimates $\hat{\Lambda}_i$ of the corresponding cumulative baseline hazards of the usual Cox model assuming linearity in V and W and consider the Poisson regression model with known offsets:

$$E(N_i|v_i,w_i)=\exp\{f(v_i)+g(w_i)\}\hat{\Lambda}_i \tag{1}$$

We shall approximate the functions $f(\cdot)$ and $g(\cdot)$ by cubic splines. Natural cubic splines with a knot at each data point are a popular choice for semi-parametric modelling [20, 21] and their optimality property of minimizing the residual sum of squares subject to a roughness penalty on the second derivative is well-known for continuous outcomes. Similar optimality properties hold for the case of generalized outcomes and have been discussed by Green and Silverman [21] in the broader context of thin-plate spline smoothing [22]. We prefer to use the radial basis representation of cubic splines (as opposed to the B-spline representation) as these have a natural extension for modelling geostatistical predictors [23]. This consists of rewriting the functions $f(\cdot)$ and $g(\cdot)$ in terms of polynomial and truncated cubic basis functions. We shall use the low-rank approximation to smoothing splines [1]. Our problem thus reduces to one of fitting the model,

$$E(N_i|\mathbf{x}_i)=\exp\left\{\beta_v v_i+\theta_v v_i^2+\sum_{k=1}^{K_v}u_k^v|v_i-\kappa_k^v|^3+\beta_w w_i+\theta_w w_i^2+\sum_{k=1}^{K_w}u_k^w|w_i-\kappa_k^w|^3\right\}\hat{\Lambda}_i \tag{2}$$

subject to penalization of the knot coefficients u_k^v and u_k^w. Here $\kappa_1^v,\kappa_2^v,\ldots,\kappa_{K_v}^v$ and κ_1^w, $\kappa_2^w,\ldots,\kappa_{K_w}^w$ are knots in the v and w directions, respectively. These are chosen as the observed sample quantiles of the v_i's and w_i's with a maximum of 20–40 knots for each. A key connection is that subjecting each set of knot coefficients u_k^v and u_k^w to quadratic penalties $\mathbf{u}_k^{v\mathrm{T}}\mathbf{G}_v^{-1}\mathbf{u}_k^v$ and $\mathbf{u}_k^{w\mathrm{T}}\mathbf{G}_w^{-1}\mathbf{u}_k^w$ is equivalent to treating them as independent normally distributed random effects from $\mathrm{N}(\mathbf{0},\mathbf{G}_v)$ and $\mathrm{N}(\mathbf{0},\mathbf{G}_w)$ populations, respectively. Specifically, if we define $\boldsymbol{\beta}=(\beta_v,\theta_v,\beta_w,\theta_w)^{\mathrm{T}}$, $\mathbf{u}=(\mathbf{u}_k^v,\mathbf{u}_k^w)^{\mathrm{T}}$, $\mathbf{X}=[v_i\ v_i^2\ w_i\ w_i^2]_{1\leqslant i\leqslant n}$ and $\mathbf{Z}=[\mathbf{Z}_v\ \mathbf{Z}_w]$, where

$$\mathbf{Z}_v=[|v_i-\kappa_k^v|^3]_{\substack{1\leqslant i\leqslant n\\ 1\leqslant k\leqslant \kappa_v}}\quad\text{and}\quad \mathbf{Z}_w=[|w_i-\kappa_k^w|^3]_{\substack{1\leqslant i\leqslant n\\ 1\leqslant k\leqslant \kappa_w}}$$

then penalized likelihood estimation for the original model in (2) is equivalent to estimation using the penalized quasi-likelihood (PQL) algorithm [24] for the following mixed model:

$$N_i|\mathbf{u}\sim\mathrm{Poisson}[\exp\{(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})_i\hat{\Lambda}_i\}],\quad \mathbf{u}\sim\mathrm{N}\left(\mathbf{0},\begin{bmatrix}\mathbf{G}_v & \mathbf{0}\\ \mathbf{0} & \mathbf{G}_w\end{bmatrix}\right) \tag{3}$$

We take $\mathbf{G}_v$ to be of the form

$$\mathbf{G}_v = \sigma_v^2(\mathbf{\Omega}_v^{-1/2})(\mathbf{\Omega}_v^{-1/2})^{\mathrm{T}} \quad \text{where } \mathbf{\Omega}_v = [|\kappa_k^v - \kappa_{k'}^v|^3]_{1 \leqslant k,k' \leqslant K_v}$$

with $\mathbf{G}_w$ and $\mathbf{\Omega}_w$ defined analogously. This choice is based on the theory of thin plate splines [13, 21, 25].

To allow fitting of (3) via standard software, we reparametrize to

$$N_i|\mathbf{u} \sim \text{Poisson}[\exp\{(\mathbf{X}\boldsymbol{\beta} + \tilde{\mathbf{Z}}\tilde{\mathbf{u}})_i\hat{\Lambda}_i\}], \quad \tilde{\mathbf{u}} \sim \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_v^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_w^2\mathbf{I} \end{bmatrix}\right) \tag{4}$$

where

$$\tilde{\mathbf{Z}} = \mathbf{Z}\begin{bmatrix} \mathbf{\Omega}_v & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_w \end{bmatrix}^{-1/2} = [\tilde{\mathbf{Z}}_v\tilde{\mathbf{Z}}_w] \quad \text{and} \quad \tilde{\mathbf{u}} = \begin{bmatrix} \tilde{\mathbf{u}}^v \\ \tilde{\mathbf{u}}^w \end{bmatrix} = \begin{bmatrix} \mathbf{\Omega}_v & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_w \end{bmatrix}^{1/2} \mathbf{u}$$

Note that our model corresponds to *low-rank* smoothing [15, 16, 26] since the number of basis functions stays fixed at $\kappa_v + \kappa_w + 2$ regardless of the sample size. For large values of n, such as commonly arises in epidemiological applications, this reduces the computational burden without sacrificing accuracy.

4. GEOSTATISTICAL EXTENSION

Suppose that the data are of the form $(T_i, C_i, \mathbf{x}_i)$, for $i = 1, 2, \ldots, n$ where (T_i, C_i) are as before and $\mathbf{x}_i \in \mathbb{R}^2, i = 1, \ldots, n$, represent n geographical locations. A general non-parametric formulation should then consider a *bivariate* model of the form

$$E(N_i|\mathbf{x}_i) = \exp\{f(\mathbf{x}_i)\}\hat{\Lambda}_i \tag{5}$$

where $f(\cdot)$ is an unspecified but sufficiently smooth bivariate function of $\mathbf{x} \in \mathbb{R}^2$. A natural extension of the methods proposed by Kammann and Wand [1] is to consider a mixed model representation of the form

$$N_i|\mathbf{u} \sim \text{Poisson}[\exp\{(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{\mathbf{x}}\mathbf{u}_{\mathbf{x}})_i\hat{\Lambda}_i\}], \quad \mathbf{u}_{\mathbf{x}} \sim \mathrm{N}(\mathbf{0}, \mathbf{G}_{\mathbf{x}}) \tag{6}$$

where $\mathbf{X} = [\mathbf{x}_i^{\mathrm{T}}]_{1 \leqslant i \leqslant n}$ and $\mathbf{Z}_{\mathbf{x}}$ contains *bivariate* radial basis functions, analogous to $\mathbf{Z}_v$ and $\mathbf{Z}_w$. A reasonable choice is

$$\mathbf{Z}_{\mathbf{x}} = [r(\|\mathbf{x}_i - \boldsymbol{\kappa}_k^{\mathbf{x}}\|)]_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant k \leqslant K^{\mathbf{x}}}}$$

where $r(x) = x^2 \log(x)$. As in the one-dimensional case (Section 3) thin plate spline theory motivates

$$\mathbf{G}_{\mathbf{x}} = \sigma_{\mathbf{x}}^2\mathbf{\Omega}_{\mathbf{x}} \quad \text{where } \mathbf{\Omega}_{\mathbf{x}} = [r(\|\boldsymbol{\kappa}_{k'}^{\mathbf{x}} - \boldsymbol{\kappa}_k^{\mathbf{x}}\|)]_{1 \leqslant k,k' \leqslant K^{\mathbf{x}}}$$

In this and in the previous section, our choice of the matrices $\mathbf{Z}$ and $\mathbf{G}$ have been dictated by the fact that we want the penalty term used in PQL estimation (Section 5) to be equivalent to the corresponding natural roughness penalty. For the univariate predictors considered in the previous section, this is a penalty on the squared second derivative while for geostatistical data, it is a roughness functional defined in terms of the second-order partial derivatives [21]. Establishing an equivalences with PQL estimation is not immediate as this does not lead to a positive definite choice of **G**. Several options have been prescribed in the literature and we use that of low-rank smoothing advocated by French *et al.* [13].

Note that the matrix **Z** turns out to depend only on radial distances from the knots. This is an attractive feature for modelling geographical data as it means that the resulting solution does not depend on the choice of co-ordinate systems.

Reparametrization to $\mathbf{Z_x}\mathbf{\Omega_x}^{-1/2}$ again facilitates implementation via standard GLMM software. The remaining choice to be made is that of the bivariate knots $\boldsymbol{\kappa}_k^{\mathbf{x}}$, $1 \leqslant k \leqslant K^{\mathbf{x}}$. An effective strategy is to use a *space filling algorithm* [10, 11]. The function `cover.design()` in the `R` package `fields` facilitates this algorithm. An alternative option is to use a clustering algorithm such as CLARA [27] which is available in the `R` package `cluster`. The number of knots were chosen based on the recommendations of Ruppert *et al.* [23] who advocate selecting $K = \max\{20, \min(n/4, 150)\}$, if the sample size does not exceed 1500 and applying the same algorithm to a random sample of size 1500 if it does.

We can combine the models in (4) and (6) to arrive at the single GLMM

$$N_i|\mathbf{u} \sim \text{Poisson}[\exp\{(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i \hat{\Lambda}_i\}], \quad \mathbf{u} \sim \text{N}(\mathbf{0}, \mathbf{G}) \tag{7}$$

Here

$$\mathbf{X} = [v_i \;\; v_i^2 \;\; w_i \;\; w_i^2 \;\; \mathbf{x}_i^{\text{T}}]_{1 \leqslant i \leqslant n}, \quad \mathbf{Z} = [\tilde{\mathbf{Z}}_v \;\; \tilde{\mathbf{Z}}_w \;\; \tilde{\mathbf{Z}}_{\mathbf{x}}] \quad \text{and} \quad \mathbf{G} = \begin{bmatrix} \sigma_v^2\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_w^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_{\mathbf{x}}^2\mathbf{I} \end{bmatrix}$$

As mentioned earlier, extension to the case where there are more than two covariates is immediate. Note also **X** does not include a column corresponding to an intercept since this is accounted for by the baseline hazard. However, predictors known to have a linear log-hazard ratio can be incorporated as additional columns of **X**.

5. INFERENCE

As mentioned in Section 1, one advantage of using the mixed model-based approach is that we are able to get likelihood-driven estimates of all parameters. The following section gives the details of the likelihood-based procedures used. We assume as before that the approximations of the baseline hazard has no effect on subsequent parameter estimation.

5.1. Asymptotic distribution

For estimation of $\boldsymbol{\beta}$ and prediction of **u**, the PQL approach can be shown to reduce to an iteratively reweighted least squares scheme with Fisher scoring updates given by

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} \longleftarrow (\mathbf{C}^{\text{T}}\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C} + \mathbf{B})^{-1}\mathbf{C}^{\text{T}}\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{N}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}} \tag{8}$$

where

$$\mathbf{C} = [\mathbf{X}, \mathbf{Z}], \quad \mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}} = \mathrm{diag}\{\exp(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}})\},$$

$$\mathbf{N}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}} + \mathbf{W}^{-1}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\{\mathbf{N} - \exp(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}}) - \hat{\Lambda}\}$$

$\mathbf{N}$ contains the N_i's and $\hat{\Lambda}$ contains the $\hat{\Lambda}_i$'s.

$$\mathbf{B} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{bmatrix}$$

Details are provided by Breslow and Clayton [24] and Wolfinger and O'Connell [28]. Approximate standard errors can be based on the above estimating equations leading to

$$\widehat{\mathrm{Cov}}\left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} \middle| \mathbf{u}\right) \simeq (\mathbf{C}^{\mathrm{T}}\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C} + \hat{\mathbf{B}})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C}(\mathbf{C}^{\mathrm{T}}\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C} + \hat{\mathbf{B}})^{-1}$$

where $\hat{\mathbf{B}}$ contains the PQL estimates of the variance components. Inference is based on the asymptotic result

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\mathbf{u}} - \mathbf{u} \end{bmatrix} \sim \mathrm{N}\left(\mathbf{0}, \left(\mathbf{C}^{\mathrm{T}}\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C} + \hat{\mathbf{B}}\right)^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C}\left(\mathbf{C}^{\mathrm{T}}\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C} + \hat{\mathbf{B}}\right)^{-1}\right) \tag{9}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ are the PQL estimates of $\boldsymbol{\beta}$ and $\mathbf{u}$, respectively. Note that the low-rank approach is crucial to the validity of this asymptotic statement since otherwise the number of parameters would grow with the sample size leading to more complicated asymptotics.

The formulation of the PQL estimation procedure as an approximate linear model on a pseudo-response led Breslow and Clayton [24] to suggest an approximate REML type criterion for estimating the variance components in $\mathbf{G}$ given by

$$-\tfrac{1}{2}\log|\mathbf{V}| - \tfrac{1}{2}\log|\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X}| - \tfrac{1}{2}(\mathbf{N}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{N}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{10}$$

where $\mathbf{V} = \mathbf{W}^{-1}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}} + \mathbf{Z}\mathbf{G}\mathbf{Z}^{\mathrm{T}}$. Specifically, fixing $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ at their current iterates, they suggest updating estimates of the variance components using ML or REML on this pseudo-data. This leads to an iterative scheme for recursively estimating $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ and the variance components in $\mathbf{G}$.

5.2. Degrees of freedom

The degrees of freedom of a fit is defined as the trace of the hat matrix defined as

$$\mathbf{H} = \mathbf{C}(\mathbf{C}^{\mathrm{T}}\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C} + \hat{\mathbf{B}})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}} \tag{11}$$

One motivation for this definition is provided by the expression for the Fisher update in equation (8) and its analogy to linear additive modelling. The degree of freedom for a predictor can be obtained by considering only the corresponding columns of the **C** matrix in (11) and provides a measure of the amount of smoothing which is comparable across various smoothing techniques.

Using a generalized linear mixed model-based approach leads to automatic likelihood-based selection of the amount of smoothing for each additive component. This is obtained by maximizing an approximate 'REML-type' criterion obtained from the iteratively reweighted least squares scheme as shown by Breslow and Clayton [24].

6. ANALYSIS OF EAST BOSTON DATA

The geostatistical model described in Section 4 was implemented using the `R` library `SemiPar` available from the `CRAN` website. The code for the analysis has been presented in Appendix A. We present the analysis using fully automated smoothing parameter choice based on REML although several choices of degrees of freedom were considered for the geographical component but did not lead to appreciably different hazard estimates. Table II summarizes the results for the linear components. Note that the number of LRI's in the first two years of life is a count variable ranging from 0 to 9 for the sampled observations. This was modelled as a categorical predictor with three categories corresponding to $0, 1\text{–}2, >2$ infections.

Figure 3 shows confidence intervals for the categorical predictors from the unadjusted (univariate) and adjusted (additive) analysis. There are some discrepancies in the two sets of confidence intervals, most notably for the effect of `lri2`. This suggests that geography acts as a confounder in the sense that adjusting for geographical variations distorts the hazard ratios for other risk factors.

Table II. Summary of final REML-based fit of geoadditive model for East Boston asthma data.

	Confidence interval
`gender`	[0.58,1.08]
`lri2`	
Baseline: (lri2 = 0)	
(0 < `lri2` < 2)	[0.99,2.01]
(`lri2` > 2)	[4.77,10.86]
`educ`	
Baseline: `educ` = less than high school	
Post-college	[0.54,1.72]
College	[0.67,1.81]
High school	[0.74,1.54]
`race`	
Baseline: `race` = Black	
Hispanic	[0.35,1.44]
White	[0.41,1.63]
`mat.asthma`	[0.92,2.24]
	df
`ciguse`	1
`cotinine`	1
Longitude, latitude	2

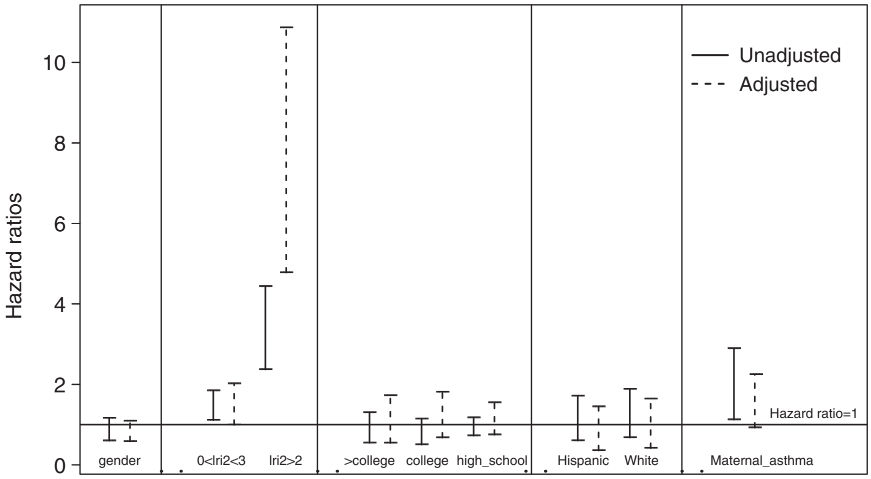

Figure 3. Adjusted and unadjusted confidence intervals for hazard ratios.

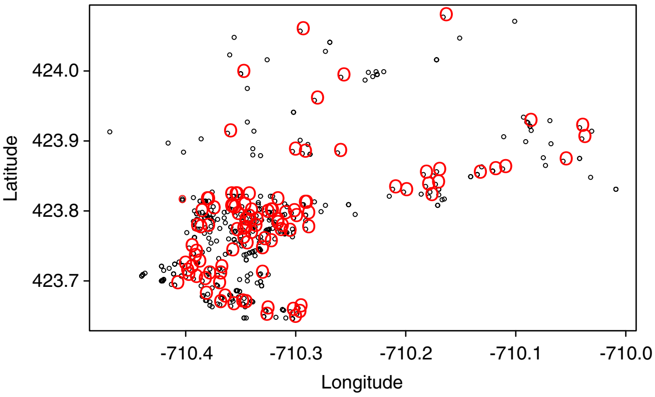

Figure 4. Knot selection for the geographical component.

REML estimates of the variance components corresponding to the continuous univariate predictors are zero meaning that we do not find statistically significant evidence of non-linearity in either of these predictors.

The geostatistical component was fit using 100 knots as per the recommendations of Ruppert *et al.* [4]. A plot showing the observed locations and the selected knots is shown in Figure 4.

Figure 5 shows an image plot of the estimated hazard function over the sampled locations from fitting the additive model. As with the continuous predictors, we do not find statistically significant evidence of residual variation in times to first asthma event although there does appear to be a gradient from South–West to North–East. This suggests that the risk of asthma

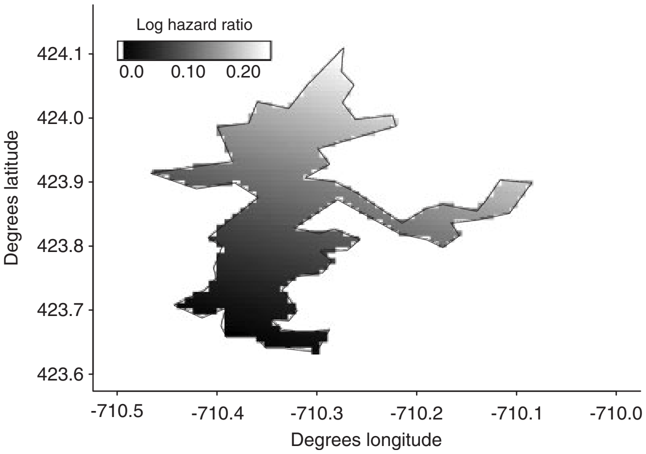

Figure 5. Contour plot of estimated hazard ratio.

is greater for children in the northern suburbs of greater Boston such as Chelsea, Winthrop and Revere than in the inner city.

7. CLOSING REMARKS

Our method performs well in recovering the true model for simulated examples of additive failure time data where the geographical component of the log-hazard ratio was obtained as linear combinations of bivariate normal densities and several choices of designs and censoring patterns were considered. However, in case of application to the motivating data set, we are faced with concerns about the level of noise in the data as well as possible sources of epidemiological bias. Our main concern is that our outcome is only the age to first physician diagnosis which may or may not correlate well with age at first onset of asthma, especially so for populations from inner-city neighbourhoods.

The application raises some further issues for future research. One such is the need to address concerns about automatic smoothing parameter selection. By reducing the model to a generalized mixed model, it is possible to address this concern by applying SiZer type techniques.

In our analysis, we restricted our attention to covariates found to be significant in single predictor fits or considered to be possible confounders based on prior knowledge. For applications where there may not be any prior knowledge or where a more automated scheme is considered desireable, an interesting possibility is that of applying model-selection criteria developed for generalized mixed models for model selection for failure time data. Model diagnostics for such data typically based on martingale residuals is complicated and a simpler alternative could be to choose from several competing models based on the value of a criterion such as the AIC.

APPENDIX A

The above analysis was carried out using the R library, SemiPar (downloadable from www.cran.r-project.org) for fitting semi-parametric regression using the mixed model approach. The code is presented below:

```
attach(asthma)

model.cox <- Surv(surv.age,surv.cens) ~ Latitude + Longitude +
              mcottot + mcigtot+
              bl + gender + (educ ==, "coll_nograd") +
              (educ == "hi_sch_grad") + (educ == "lt_hi_sch") +
              (educ == "post_coll") + (educ == "tech_sch") +
              ((0<lri2)&(lri2<2)) + (lri2>2)+
              (race == "hispanic") +
              (race == "white") + mevasth

 library(survival)

 asthfit<-coxph(model.cox)
 exp.fit<-predict(asthfit, type = "expected")
 xbeta<-predict(asthfit, type = "lp")
 newtime<-exp(-xbeta)*exp.fit # estimate of expected baseline hazard

# Fit additive model using SemiPar

model.glmm <- surv.cens ~  Latitude + Longitude + mcottot + mcigtot+
              bl + gender + (educ == "coll_nograd") +
              (educ == "hi_sch_grad") + (educ == "lt_hi_sch") +
              (educ == "post_coll") + (educ == "tech_sch") +
              ((0<lri2)&(lri2<2)) + (lri2>2)+
              (race == "hispanic") +
              (race == "white") + mevasth +
              f(mcigtot) + f(mcottot) +
              f(Longitude, Latitude, degree = 5)+
          offset(log(newtime))

library(SemiPar)

spm.fit <- spm(model.glmm,family = "poisson")
```

ACKNOWLEDGEMENTS

This research was supported by U.S. National Institutes of Health grant R01-ES10844-01. We are grateful to Dr Rosalind Wright and Ms Marina Jacobson for providing the data and advice about the application.

REFERENCES

1. Kammann EE, Wand MP. Geoadditive models. *Applied Statistics* 2003; **52**:1–18.
2. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994; **81**:515–526.
3. Grambsch PM, Therneau TM, Fleming TR. Diagnostic plots to reveal functional form for covariates in multiplicative intensity models. *Biometrics* 1995; **51**:1469–1482.
4. Therneau TM, Grambsch PM. *Modeling Survival Data*: *Extending the Cox Model*. Springer: New York, 2000.
5. Therneau TM, Grambsch PM. Martingale-based residuals for survival models. *Biometrika* 1990; **77**:147–160.
6. Gray RJ. Spline-based tests in survival analysis. *Biometrics* 1994; **50**:640–652.
7. Gray RJ. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association* 1992; **87**:942–951.
8. Li Y, Ryan LM. Modeling spatial survival data using semiparametric frailty models. *Biometrics* 2003; **58**: 287–297.
9. Banerjee S, Carlin BP. Semiparametric spatio-temporal frailty models. *Environmetrics* 2003; **14**:523–535.
10. Nychka DW. Spatial process estimates as smoothers. In *Smoothing and Regression*, Schimek M (ed.). Springer: Heidelberg, 2000.
11. Nychka D, Saltzman N. Design of air quality monitoring networks. In *Case Studies in Environmental Statistics*, Nychka D, Cox L, Piegorsch W (eds), Lecture Notes in Statistics. Springer: New York, 1998.
12. O'Connell MA, Wolfinger RD. Spatial regression models, response surfaces, and process optimization. *Journal of Computational and Graphical Statistics* 1997; **6**:224–241.
13. French JL, Kammann EE, Wand MP. Comment on paper by Ke and Wang. *Journal of the American Statistical Association* 2001; **96**:1285–1288.
14. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science* 1996; **89**:89–121.
15. Hastie TJ. Pseudosplines. *Journal of the Royal Statistical Society*, *Series B* 1996; **58**:379–396.
16. Hastie TJ, Tibshirani R. *Generalized Additive Models*. Chapman & Hall: London, 1996.
17. Carr W, Zeitel L, Weiss K. Variations in asthma hospitalizations and deaths in New York city. *American Journal of Public Health* 1992; **82**:59–65.
18. Marder D, Targonsky P, Orris P, Persky V, Addington W. Effect of racial and socioeconomic factors on asthma mortality in Chicago. *Chest* 1992; **101**:79S–83S.
19. Anderson PK. Testing goodness of fit of Cox's regression and life model. *Biometrics* 1982; **38**:67–77.
20. Eubank RL. *Spline Smoothing and Nonparametric Regression*. Marcel Dekker: New York, 1988.
21. Green PJ, Silverman BW. *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall: London, 1994.
22. Wood SN. Thin plate regression splines. *Journal of the Royal Statistical Society*, *Series B* 2003; **65**:95–114.
23. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge University Press: New York, 2003.
24. Breslow NE, Clayton DG. Approximated inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**:9–25.
25. Wahba G. *Spline Models for Observational Data*. SIAM: Philadelphia, 1990.
26. Cai T, Hyndman RJ, Wand MP. Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics* 2002; **11**:784–798.
27. Struyf A, Hubert M, Rousseeuw PJ. *Journal of Statistical Software* 1996; **1**:1–30.
28. Wolfinger R, O'Connell M. Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* 1993; **48**:233–243.
29. Adebayo SM, Fahrmeir L. Analysing child mortality in Nigeria with geoadditive discrete-time survival models. *Statistics in Medicine* 2005; **24**:709–728.
30. Chaudhuri P, Marron JS. SiZer for exploration of structures in curves. *Journal of the American Statistical Association* 1999; **94**:807–823.
31. Cressie N. *Statistics for Spatial Data*. Wiley: New York, 1993.
32. Ganguli B, Wand MP. Feature significance in generalized additive models. Unpublished manuscript, 1994.