*Biostatistics* (2004), **5**, 2, *pp.* 177–191
*Printed in Great Britain*

# Generalized additive models for cancer mapping with incomplete covariates

JONATHAN L. FRENCH[†]
*Biostatistics, Global Research and Development, Pfizer, Inc., 50 Pequot Avenue, New London, CT, 06320, USA,*
Jonathan_L_French@groton.pfizer.com

MATTHEW P. WAND
*Harvard School of Public Health, Boston, USA*

SUMMARY

Maps depicting cancer incidence rates have become useful tools in public health research, giving valuable information about the spatial variation in rates of disease. Typically, these maps are generated using count data aggregated over areas such as counties or census blocks. However, with the proliferation of geographic information systems and related databases, it is becoming easier to obtain exact spatial locations for the cancer cases and suitable control subjects. The use of such point data allows us to adjust for individual-level covariates, such as age and smoking status, when estimating the spatial variation in disease risk. Unfortunately, such covariate information is often subject to missingness. We propose a method for mapping cancer risk when covariates are not completely observed. We model these data using a logistic generalized additive model. Estimates of the linear and non-linear effects are obtained using a mixed effects model representation. We develop an EM algorithm to account for missing data and the random effects. Since the expectation step involves an intractable integral, we estimate the E-step with a Laplace approximation. This framework provides a general method for handling missing covariate values when fitting generalized additive models. We illustrate our method through an analysis of cancer incidence data from Cape Cod, Massachusetts. These analyses demonstrate that standard complete-case methods can yield biased estimates of the spatial variation of cancer risk.

*Keywords*: Binary response; Expectation Maximization; Generalized linear model; Laplace approximation; Logistic regression; Method of weights; Missing data.

## 1. INTRODUCTION

Geographical information systems are becoming widely used in environmental health and epidemiology. Correspondingly, investigators would like to produce summary maps of their data. For example, in the context of environmental epidemiology, we might be interested in a map of relative cancer incidence rates or cancer risk. Many studies, however, are unable to collect complete covariate information for each subject. These values may be either missing by design (e.g. when some data are collected on only a subset of subjects) or missing unintentionally. In this paper we propose a method to estimate maps of covariate adjusted relative cancer risk in the presence of incomplete covariate data.

[†]To whom correspondence should be addressed.

This work was motivated by a study of cancer incidence on Cape Cod, Massachusetts, USA. For nearly twenty years the Massachusetts Department of Public Health (MDPH) has maintained a cancer registry database which records incident cases for 22 types of cancers, including lung, breast, and prostate cancers. The report summarizing these data for the period 1982–92 showed elevated standardized incidence ratios for several types of cancer in the towns of Barnstable, Falmouth, Sandwich, Bourne, and Mashpee (Massachusetts Department of Public Health, 1997). Collectively, these towns comprise a region known as Upper Cape Cod. Prostate cancer in men was of particular concern since it appeared to be the most elevated. The MDPH was interested in determining if the elevated cancer rates were due to environmental factors, including contaminants at several sites on the Massachusetts Military Reservation (MMR) that have been listed on the US Environmental Protection Agency's National Priority List. These sites are of significant interest because the MMR is situated above the sole-source aquifer for the Upper Cape.

In addition to the type of cancer, the registry maintains data on the patient's residence and age at the time of diagnosis, as well as gender and smoking status. The data are routinely summarized by gender at the town level, but raw, geo-coded residence locations have been collected for this study. Assuming that adverse environmental effects are likely to cause only certain types of cancer, this enables us to overcome the problem of adjusting for population density which is inherent in calculating standardized incidence ratios. Such data are available only at the census tract level which is much coarser than the cancer data. This type of analysis also allows us to adjust for individual level covariates such as age and smoking status.

In this paper we analyze data for incidence of prostate cancer on Upper Cape Cod. Data have been compiled for $n = 3191$ men diagnosed with one of the 22 types of cancer during the period 1987–94. While age and location data were collected for all patients, smoking status was not obtained for approximately 29% of the men. The upper-left panel of Figure 1 displays the approximate locations of the residences of the patients. For confidentiality reasons, the points have been jittered. The solid and empty dots correspond to prostate cancer and other cancer cases, respectively. The lines correspond to the census tract boundaries. We see that there are several areas with a high density of cancer cases; these correspond to the population centers on the Upper Cape and are not necessarily associated with any sort of exposures. By using relative cancer mapping, we assume that all cancer cases provide a reasonable surrogate for population density across the region. Potential difficulties with relative cancer mapping are discussed in Section 4.

Table 1 displays summary statistics for age and smoking status classified by cancer type. Age is a categorical variable with categories defined by age decade (i.e. age = 1 for ages between 1 and 10 years, age = 2 for 11–20 years, etc.). Smoking status has two categories: ever-smokers (smoke = 1) and never-smokers (smoke = 0). Of the cohort of 3191 men, 998 (31.3%) were prostate cancer patients. We see that prostate cancer patients are less likely to be smokers than are patients with other types of cancer. This is partially explained by the relatively large number of lung cancer patients, most of which were smokers. We also see that the pattern of missing data depends on cancer type. Smoking status is missing for 41.0% of the prostate cancer patients, while 24.2% of the other cancer patients did not provide information about smoking status. Thus, these data are clearly not missing completely-at-random in the sense of Little and Rubin (1987).

Fitting maps of relative cancer risk, and bivariate surfaces in general, can be accomplished in a number of ways. The most commonly used methods are kriging (e.g. Cressie, 1993) and thin plate splines (Wahba, 1990). The extension of kriging to discrete response data is not straightforward. Methods to accomplish this, such as indicator kriging, have been developed, but are somewhat ad hoc. Diggle *et al.* (1998) have recently proposed an attractive method for handling such data using Markov chain Monte Carlo methods. We take a different approach, using generalized additive models with a mixed model representation.

Generalized additive models (e.g. Hastie and Tibshirani, 1990) provide a flexible means of handling non-linear covariate effects. These models have gained widespread popularity in a diverse set of

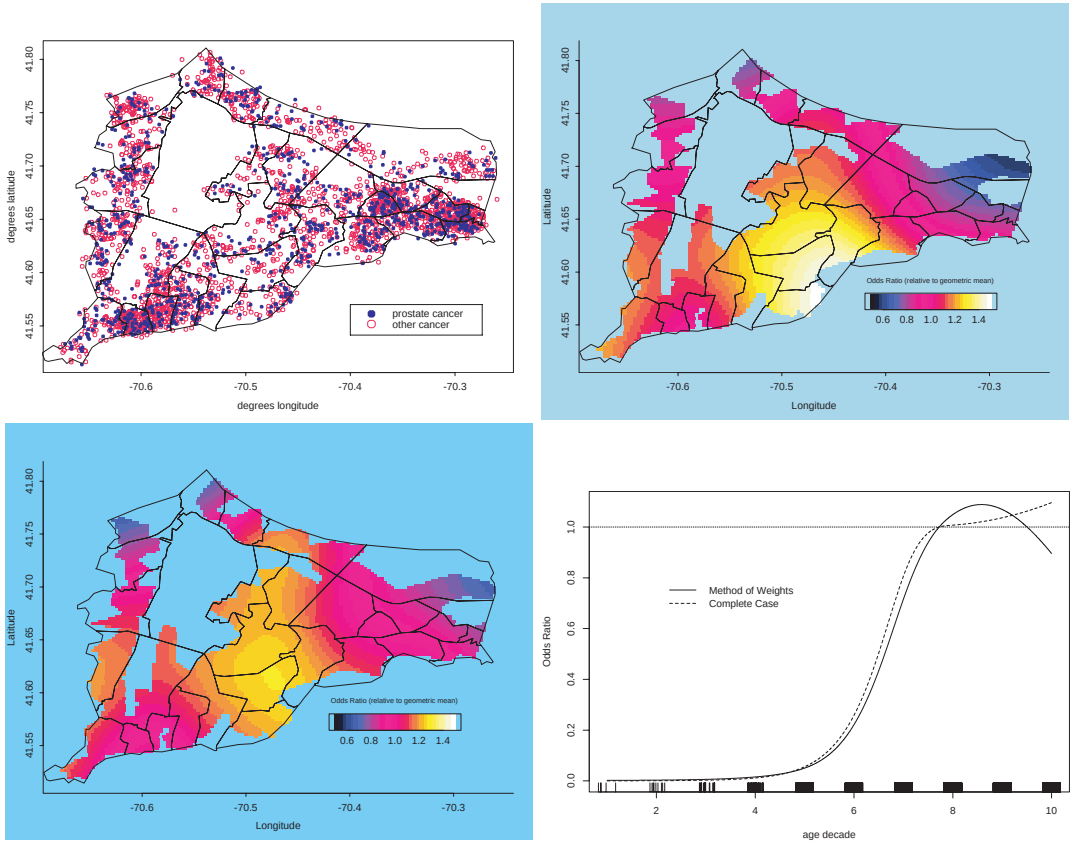


Fig. 1. Upper Cape Cod cancer data. Upper-left panel: Raw data. For confidentiality reasons, the points have been jittered. Upper-right: Complete-case estimate of the odds ratio as a function of geography. Lower-left: Method of weights estimate of the odds ratio as a function of geography. Lower-right: Estimates of the odds ratio for age relative to the average age in the cohort.

Table 1. *Summary statistics for age and smoking status, classified by cancer type for the Upper Cape Cod data*

| | Prostate cancer | | | | Other cancers | | | |
|---|---|---|---|---|---|---|---|---|
| | $n^{\text{obs}}$ | mean | sd | $n^{\text{miss}}$ (%) | $n^{\text{obs}}$ | mean | sd | $n^{\text{miss}}$ (%) |
| smoke | 588 | 0.56 | 0.50 | 410(41.0) | 1663 | 0.75 | 0.44 | 530(24.2) |
| age | 998 | 8.25 | 0.91 | 0(0.0) | 2193 | 7.76 | 1.39 | 0(0.0) |

disciplines including environmental epidemiology, environmental science, ecology, public health, political science, and economics (e.g. Linton and Härdle, 1996; Schwartz, 1997; Beck and Jackman, 1998; Davis and Speckman, 1999; Engels *et al.*, 1999; Roland *et al.*, 2000). Recently, Kammann and Wand (2003) showed how kriging could be incorporated into a generalized additive model, with representation as a generalized linear mixed model. Because mixed models can be fit using maximum likelihood, this allows us to use likelihood-based methods for estimation and handling of missing data. We develop a method for handling missing covariate values in generalized additive models using the EM algorithm (Dempster *et*

*al.*, 1977). For models other than the Gaussian response, the E-step involves an intractable integral. As in Steele (1996) we use Laplace's method (Tierney *et al.*, 1989) to approximate this expectation.

Section 2 introduces the generalized additive model, estimation of penalized regression splines via a mixed effects model, and methods for estimation in generalized linear mixed effects models with complete data. Section 3 develops an algorithm for estimation of generalized additive models in which some covariates exhibit missingness. In Section 4, we analyze the Cape Cod cancer data. We conclude with a brief discussion in Section 5. Additional computational details are provided in an appendix on Biostatistics Online.

## 2. MAPPING CANCER RISK USING GENERALIZED ADDITIVE MODELS

Suppose first that the data are $(\mathbf{x}_i, y_i)$, $1 \leqslant i \leqslant n$, where the $y_i$ are binary and $\mathbf{x}_i \in \mathbb{R}^2$ represents geographical location. To assess spatial variation in data of this type, Diggle *et al.* (1998) propose the kriging-type model

$$\text{logit}\{P(y_i = 1|S(\mathbf{x}_i))\} = \beta_0 + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i + S(\mathbf{x}_i)\,, \tag{2.1}$$

where $\{S(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^2\}$ is a stationary zero-mean stochastic process.

The right-hand side of (2.1) is the logarithm of the odds for the event $\{y_i = 1\}$, given $\mathbf{x}_i$, which we denote by $\text{LO}(\mathbf{x}_i)$. It is often useful to map an estimate of $\text{LO}(\mathbf{x}_0)$ over a mesh of $\mathbf{x}_0 \in \mathbb{R}^2$ values. This involves the bivariate function

$$\widehat{\text{LO}}(\mathbf{x}_0) = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{x}_0 + \hat{S}(\mathbf{x}_0)\,, \tag{2.2}$$

where $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}_1$ are maximum likelihood estimates and

$$\hat{S}(\mathbf{x}_0) = E\{S(\mathbf{x}_0)|\mathbf{y}\}$$

is an 'optimal' predictor of $S(\mathbf{x}_0)$.

To make (2.2) practical we require a parsimonious model for the inter-point covariances $\text{cov}\{S(\mathbf{x}), S(\mathbf{x}')\}$, $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^2$. We propose to use the isotropic model

$$\text{cov}\{S(\mathbf{x}), S(\mathbf{x}')\} = C_{\boldsymbol{\theta}}(\|\mathbf{x} - \mathbf{x}'\|)\,, \tag{2.3}$$

where $\|\mathbf{v}\| = \sqrt{\mathbf{v}^{\mathrm{T}}\mathbf{v}}$ and $C_{\boldsymbol{\theta}}$ is member of the Matérn family of covariance functions (see Stein, 1999). We will work in the subfamily corresponding to a single value of the Matérn smoothness parameter:

$$C_{\boldsymbol{\theta}}(r) = \sigma_{\mathbf{x}}^2(1 + |r|/\rho)\mathrm{e}^{-|r|/\rho}. \tag{2.4}$$

The function given in (2.4) is the simplest member of the Matérn family that yields differentiable surface estimate. Kammann and Wand (2003) provide further discussion of the choice of the Matérn smoothness parameter.

To ensure scale invariance, increase numerical stability and reduce the computational burden, we choose the range parameter $\rho$ via the simple rule

$$\hat{\rho} = \max_{1 \leqslant i,\, j \leqslant n} \|\mathbf{x}_i - \mathbf{x}_j\|. \tag{2.5}$$

Fixing $\rho$ a priori allows the use of the generalized linear model for fitting. In our experience, the smoothness of the estimate surface depends on the choice of $\rho$, but the scale of the final map is rather insensitive.

Fitting of the Diggle *et al.* (1998) model requires $n \times n$ matrix storage and inversion. Since the Upper Cape Cod cancer data set involves $n = 3191$ observations, some modification is required for practical use. An attractive solution is to use *reduced knot* or *low-rank* kriging as proposed by Nychka *et al.* (1998). Let $\{\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_{K_\mathbf{x}}\}$ be a representative subset of $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ which we will refer to as *knots*. This subset can be obtained via an efficient space filling algorithm (e.g. Johnson *et al.*, 1990; Nychka and Saltzman, 1998). Let

$$\mathbf{X} = [1\, \mathbf{x}_i \mathrm{T}]_{1 \leqslant i \leqslant n}, \quad \mathbf{Z} = [C_0(\|\mathbf{x}_i - \boldsymbol{\kappa}_k\|/\rho)]_{1 \leqslant i \leqslant n, 1 \leqslant k \leqslant K_\mathbf{x}}$$

and

$$\boldsymbol{\Omega} = [C_0(\|\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'}\|/\rho)]_{1 \leqslant k,k' \leqslant K_\mathbf{x}},$$

where $C_0(r) = (1 + |r|)e^{-|r|}$. Then low-rank kriging corresponds to fitting the logistic mixed model

$$\text{logit}\{P(y_i = 1|\mathbf{b})\} = (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b})_i, \tag{2.6}$$

where $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$, $E(\mathbf{b}) = \mathbf{0}$ and $\text{cov}(\mathbf{b}) = \sigma_\mathbf{x}^2 \boldsymbol{\Omega}^{-1}$. Typically, $\mathbf{b}$ is taken to have a Gaussian distribution. The reparametrization

$$\mathbf{Z}_* = \mathbf{Z}\boldsymbol{\Omega}^{-1/2}, \quad \mathbf{b}_* = \boldsymbol{\Omega}^{1/2}\mathbf{b}$$

means that $\text{cov}(\mathbf{b}_*) = \sigma_\mathbf{x}^2 \mathbf{I}$ and standard mixed model software, such as the `GLIMMIX` macro in `SAS`, can be used for fitting the model. The estimated log odds at $\mathbf{x}_0$ is then

$$\widehat{\text{LO}}(\mathbf{x}_0) = \mathbf{X}(\mathbf{x}_0)\hat{\boldsymbol{\beta}} + \mathbf{Z}_*(\mathbf{x}_0)\hat{\mathbf{b}}_*,$$

where

$$\mathbf{X}(\mathbf{x}_0) = [1\, \mathbf{x}_0^{\mathrm{T}}]$$

and

$$\mathbf{Z}_*(\mathbf{x}_0) = [C_0(\|\mathbf{x}_0 - \boldsymbol{\kappa}_k\|/\rho)]_{1 \leqslant k \leqslant K_\mathbf{x}}\, \boldsymbol{\Omega}^{-1/2}.$$

In the Upper Cape Cod cancer study there are data on smoking and age for which we would like to control. Since the age effect is possibly non-linear we propose to model it as an arbitrary smooth function, $f$. Assuming additivity in the logit scale, we arrive at

$$\text{logit}\{P(y_i = 1|\mathbf{b})\} = (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b})_i + \beta_s s_i + f(a_i),$$

where $a_i$ is the age of person $i$ and $s_i = 1$ if person $i$ was ever a smoker and zero otherwise.

The $\beta_s s_i$ term can be incorporated into the $\mathbf{X}\boldsymbol{\beta}$ component. There are a number of mixed model representations of smoothing that can be used to subsume the $f(a_i)$ into the generalized linear mixed model as well (Brumback and Rice, 1998; Wang, 1998; Brumback *et al.*, 1999; Lin and Zhang, 1999, Verbyla *et al.*, 1999). For example, Kammann and Wand (2003) use linear splines to achieve this. An alternative that we consider here is to simply apply the same principle for the geographic 'smooth' to the age variable. Let $\kappa_1^a, \ldots, \kappa_{K_a}^a$ be a set of knots equally spaced with respect to the quantiles of the $a_i$ and set

$$\boldsymbol{\Omega}_a = [C_0(|\kappa_k^a - \kappa_{k'}^a|/\rho_a)]_{1 \leqslant k,k' \leqslant K_a}$$

and

$$\mathbf{Z}_a = [C_0(|a_i - \kappa_k^a|/\rho_a)]_{1 \leqslant i \leqslant n, 1 \leqslant k \leqslant K_a},$$

for some $\rho_a > 0$. We use $\rho_a = \max(a_i) - \min(a_i)$.

If we redefine

$$\mathbf{X} = [1\ s_i\ a_i\ \mathbf{x}_i^{\mathrm{T}}]_{1\leqslant i\leqslant n} \tag{2.7}$$

and

$$\mathbf{Z} = [\mathbf{Z}_a\boldsymbol{\Omega}_a^{-1/2}|\mathbf{Z}_*]\ , \tag{2.8}$$

then the model has the representation

$$\text{logit}\{P(y_i = 1|\mathbf{b})\} = (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b})_i\ , \tag{2.9}$$

$$\text{Cov}(\mathbf{b}) = \text{diag}(\sigma_a^2\mathbf{1}_{K_a}, \sigma_{\mathbf{x}}^2\mathbf{1}_{K_{\mathbf{x}}})\ , \tag{2.10}$$

where $\mathbf{1}_n$ is the $1 \times n$ vector of ones.

A common convention in additive modelling is to centre the curve estimates about their means. The components of the additive model can be interpreted as effects about the mean. The same convention could be applied to the surface estimate in the kriging component of the geoadditive model. Operationally we set $\mathbf{C} = [\mathbf{X}|\mathbf{Z}]$ and let $\mathbf{C} = [\mathbf{1}|\mathbf{C}_r]$ be a partition of $\mathbf{C}$ into the intercept column and the remainder. We then work with

$$\bar{\mathbf{C}} = [\mathbf{1}|(\mathbf{I} - \tfrac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\mathbf{C}_r] \tag{2.11}$$

rather than $\mathbf{C}$. This convention is adopted in our analyses in Section 4. A consequence of this centering is that a map of $\exp\{\hat{S}(\mathbf{x}_0)\}$ over a mesh of $\mathbf{x}_0$ values plots the ratio of the odds for the event $\{y = 1\}$ at $\mathbf{x}_0$ relative to the geometric mean odds over the mapped region.

### 2.1 *Fitting as a generalized linear mixed model*

In (2.9) and (2.10) we showed that the GAM has a logistic mixed model representation. Assuming that the elements of $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ are conditionally independent given $\mathbf{b}$, the joint density of $\mathbf{y}$ and $\mathbf{b}$ is given by $p(\mathbf{y}, \mathbf{b}; \psi) = p(\mathbf{y}|\mathbf{b}; \boldsymbol{\beta})\ p(\mathbf{b}; \boldsymbol{\theta})$, where

$$p(\mathbf{y}|\mathbf{b}; \boldsymbol{\beta}) = \exp\{\mathbf{y}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}) - \mathbf{1}^{\mathrm{T}}\mathrm{B}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}) + \mathbf{1}^{\mathrm{T}}\mathrm{C}(\mathbf{y})\}\ , \tag{2.12}$$

$\mathrm{B}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}) = [\log[1 + \exp\{(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b})_i\}]]_{1\leqslant i\leqslant n}$ , $\mathrm{C}(\mathbf{y}) = \mathbf{0}$, and, assuming $\mathbf{b}$ follows a Gaussian distribution with mean zero,

$$p(\mathbf{b}; \boldsymbol{\theta}) = (2\pi)^{-M/2}\ |\mathbf{D}_{\boldsymbol{\theta}}|^{-1/2}\exp(-\tfrac{1}{2}\mathbf{b}^{\mathrm{T}}\mathbf{D}_{\boldsymbol{\theta}}^{-1}\mathbf{b})\ , \tag{2.13}$$

where $M = K_a + K_{\mathbf{x}}$ and $\mathbf{D}_{\boldsymbol{\theta}}$ is given by (2.10). Let $\psi \equiv (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})^{\mathrm{T}}$. We refer to $p(\mathbf{y}, \mathbf{b}; \psi)$ as both the complete-data density and complete-data likelihood.

Maximum likelihood inference for generalized linear mixed effects models is based on maximizing the observed-data likelihood

$$\begin{aligned} L(\psi; \mathbf{y}) &= \int_{\mathbb{R}^M} p(\mathbf{y}, \mathbf{b}; \psi)\, \mathrm{d}\mathbf{b} \\ &= (2\pi)^{-M/2}\ |\mathbf{D}_{\boldsymbol{\theta}}|^{-1/2}\exp\{\mathbf{1}^{\mathrm{T}}\mathrm{C}(\mathbf{y})\}I(\boldsymbol{\beta}, \boldsymbol{\theta})\ , \end{aligned} \tag{2.14}$$

where

$$I(\boldsymbol{\beta}, \boldsymbol{\theta}) = \int_{\mathbb{R}^M} \exp\{\mathbf{y}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}) - \mathbf{1}^{\mathrm{T}}\mathrm{B}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}) - \tfrac{1}{2}\mathbf{b}^{\mathrm{T}}\mathbf{D}_{\boldsymbol{\theta}}^{-1}\mathbf{b}\}\, \mathrm{d}\mathbf{b}. \tag{2.15}$$

In most cases, including the logistic mixed model used in (2.9) and (2.10), this integral is intractable. Thus, we cannot find maximum likelihood estimates using the exact observed-data likelihood.

The `GLIMMIX` macro in `SAS` overcomes this with an algorithm commonly referred to as penalized quasi-likelihood (PQL) that essentially replaces (2.15) by a Laplace approximation. Another approach to finding maximum likelihood estimate of $\psi$ is to use the EM algorithm (Dempster *et al.*, 1977). The EM algorithm is a general purpose algorithm for finding the mode of a likelihood or posterior density function. Viewing the random effects as latent variables, the EM algorithm iterates between calculating the conditional expectation of the complete-data log likelihood, $\ell(\psi; \mathbf{y}, \mathbf{b}) = \log p(\mathbf{y}, \mathbf{b}; \psi)$, given the observed data and maximizing this expected value as a function of $\psi$. Dempster *et al.* (1977) have shown that the EM algorithm will lead to the maximum likelihood estimates based on the observed-data likelihood given in equation (2.14). In fact, since the observed-data likelihood is log-concave, the EM algorithm will work quite well because it will not get stuck in local modes. In the context of the generalized linear mixed effects model, the $(m+1)$th iteration of the EM algorithm is

**E-step:** Calculate $Q(\psi|\psi^{(m)}) = E\{\ell(\psi; \mathbf{y}, \mathbf{b})|\mathbf{y}, \psi^{(m)}\}$, where the expectation is taken with respect to $p(\mathbf{b}|\mathbf{y}; \psi^{(m)})$.

**M-Step:** Set $\psi^{(m+1)} = \text{argmax}_{\psi}\, Q(\psi|\psi^{(m)})$.

The algorithm alternates between these two steps until there is a sufficiently small change in the parameter values between iterations.

When working with generalized linear mixed effects models, the expectation step involves evaluating the quantity

$$\begin{aligned} Q(\psi|\psi') &= \int \ell(\psi; \mathbf{y}, \mathbf{b})\ p(\mathbf{b}|\mathbf{y}; \psi')\, \mathrm{d}\mathbf{b} \\ &= \frac{\int \ell(\psi; \mathbf{y}, \mathbf{b})\ p(\mathbf{y}, \mathbf{b}; \psi')\, \mathrm{d}\mathbf{b}}{\int p(\mathbf{y}, \mathbf{b}; \psi')\, \mathrm{d}\mathbf{b}}, \end{aligned} \tag{2.16}$$

which is intractable for most models. To alleviate this problem, several authors have use a Monte Carlo EM algorithm (Wei and Tanner, 1990) in which (2.16) is replaced by a Monte Carlo approximation based on a sample from $p(\mathbf{b}|\mathbf{y}; \psi)$. McCulloch (1997) and Ibrahim *et al.* (2001) propose obtaining the Monte Carlo sample using Markov chain methods, such as the Metropolis–Hastings algorithm or Gibbs sampling. Since Markov chain sampling induces dependence, both Walker (1996) and Booth and Hobert (1999) suggest alternative approaches to obtaining the Monte Carlo sample. The methods of Booth and Hobert (1999) and Ibrahim *et al.* (2001) also provide algorithms for adaptively choosing the Monte Carlo sample size.

The Monte Carlo EM methodology is quite computationally intensive and more suited to smaller numbers of random effects and sample sizes. Neither is the case for the geoadditive model for the Upper Cape Cod data where $\mathbf{b}$ may have dimension of 100 and $n = 3191$. The MCEM approach is simply not practical and an asymptotic approximation seems necessary. Steele (1996) describes an EM algorithm that alternates between calculating $\widehat{\mathsf{D}}_{\psi} Q(\psi|\psi')$, a second-order Laplace approximation to $\mathsf{D}_{\psi} Q(\psi|\psi')$, and solving $\widehat{\mathsf{D}}_{\psi} Q(\psi|\psi') = \mathbf{0}$. Assuming that the order of differentiation and integration can be interchanged, a standard assumption of the EM algorithm,

$$\mathsf{D}_{\psi} Q(\psi|\psi') = \int \mathsf{D}_{\psi} \ell(\psi; \mathbf{y}, \mathbf{b})\ p(\mathbf{b}|\mathbf{y}; \psi')\, \mathrm{d}\mathbf{b}. \tag{2.17}$$

Regularity conditions allow the fully exponential Laplace approximation (Tierney *et al.*, 1989) to the expected complete-data score vector (2.17), but not the expected complete-data log-likelihood (Steele,

1996). The Laplace approximation approximation to $\mathsf{D}_\psi Q(\psi|\psi')$ is given by

$$\widehat{\mathsf{D}}_\psi Q(\psi|\psi') = \{\mathsf{D}_\psi \ell(\psi; \mathbf{y}, \mathbf{b}) + C(\psi; \mathbf{y}, \mathbf{b})\}\big|_{\mathbf{b}=\hat{\mathbf{b}}(\psi')}, \tag{2.18}$$

where

$$\hat{\mathbf{b}}(\psi) \equiv \text{argmax}_{\mathbf{b}}\, p(\mathbf{y}, \mathbf{b}; \psi).$$

The term $C(\psi; \mathbf{y}, \mathbf{b}) = \{C(\psi_k; \mathbf{y}, \mathbf{b})\}_{1 \leqslant k \leqslant p+S}$ is an adjustment that allows $\widehat{\mathsf{D}}_\psi Q(\psi|\psi')$ to differ from $\mathsf{D}_\psi Q(\psi|\psi')$ by a $O(n^{-2})$ term. The individual correction terms are given by

$$C(\psi_k; \mathbf{y}, \mathbf{b}) = \tfrac{1}{2}\text{tr}\left[\mathbf{A}^{-1}\left\{\tfrac{\partial}{\partial\psi_k}\mathbf{A} - \mathbf{Z}^{\mathrm{T}}\mathbf{U}\,\text{diag}\left(\mathbf{Z}\mathbf{A}^{-1}\left[\mathsf{D}_{\mathbf{b}}\left\{\tfrac{\partial}{\partial\psi_k}\ell\,(\psi; \mathbf{y}, \mathbf{b})\right\}\right]^{\mathrm{T}}\right)\mathbf{Z}\right\}\right], \tag{2.19}$$

where $\mathbf{U} = \text{diag}\{\mathrm{B}'''(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b})\}$ and $\mathbf{A} = -\mathsf{H}_{\mathbf{b}}\ell(\psi; \mathbf{y}, \mathbf{b}) = \mathbf{Z}^{\mathrm{T}}\,\text{diag}\{\mathrm{B}''(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b})\}\,\mathbf{Z} + \mathbf{D}_{\boldsymbol{\theta}}^{-1}$.

Steele (1996) notes that a useful first-order approximation to $\mathsf{D}_\psi Q(\psi|\psi')$ can be obtained by ignoring the last term in (2.18). Using this simplified approximation, the estimating equations for $\boldsymbol{\beta}$ in the Laplace EM algorithm (Steele, 1996) and the PQL algorithm (Breslow and Clayton, 1993) are identical. That is, both algorithms solve $\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \widehat{\boldsymbol{\mu}}) = \mathbf{0}$, where $\widehat{\boldsymbol{\mu}} = \mathrm{B}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\hat{\mathbf{b}})$. The two algorithms differ in the manner in which $\boldsymbol{\theta}$ is estimated. However, when the estimates of $\boldsymbol{\theta}$ are similar, the Laplace EM algorithm should yield more accurate estimates of the fixed effects since they are based on a second-order approximation rather than a first-order approximation (Steele, 1996).

## 3. MAPPING CANCER RISK WITH INCOMPLETE COVARIATES

Missing covariate values are common in environmental and public health data. These values may be missing by design (e.g. when some data are collected only on a subset of subjects) or, as is more commonly the case, missing unintentionally. The standard approach is to perform a complete-case analysis, in which cases with missing values are excluded from the analysis. This approach can lead to a substantial loss of information if many variables exhibit missingness. Worse still, it can yield biased parameter estimates if the missingness depends on the response variable (Jones, 1996). Likelihood-based methods, such as the EM algorithm, use all of the available data, including those observations with missing covariate values. They are particularly attractive because they provide valid inference when the missing values are missing-at-random in the sense of Little and Rubin (1987) and can be extended to include situations when the data are not missing-at-random. One interesting feature of the complete-case analysis is that estimates are not biased when the missingness depends on the missing covariate values, a property that is not shared by likelihood-based methods.

In the Upper Cape Cod cancer study, the geographic location and age variables are completely observed for all subjects, but smoking status is missing for 29% of the subjects. Table 1 shows that the missingness in smoking status depends on the cancer type, with smoking status missing for 41.0% of prostate cancer patients and 24.2% of patients with other types of cancer. This is precisely the situation in which complete case methods will lead to biased inference. So, an alternative method is needed.

We begin by considering a model with $D$ discrete and $S$ continuous covariates, whose values for subject $i$ will be denoted by $\mathbf{d}_i$ and $\mathbf{c}_i$, respectively. To simplify the notation, we assume that we are smoothing each of the $S$ continuous covariates. Suppose that the response $\mathbf{y}$ and all continuous covariates are completely observed, but that some of the values for $\mathbf{d}_i$ may be missing-at-random for each subject. For the $i$th subject, let $\mathbf{d}_i^{\text{obs}}$ and $\mathbf{d}_i^{\text{miss}}$ denote the vectors of observed and missing values and $n_i$ denote the number of possible combinations of values for $\mathbf{d}_i^{\text{miss}}$. For the appropriate $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{b}$, we can express the complete-data density as

$$p(\mathbf{y}, \mathbf{d}, \mathbf{b}|\mathbf{c}; \phi) = p(\mathbf{y}|\mathbf{b}, \mathbf{c}, \mathbf{d}; \beta)\; p(\mathbf{d}|\mathbf{c}; \gamma)\; p(\mathbf{b}; \theta)\;,$$

where $\phi = (\beta^{\mathrm{T}}, \gamma^{\mathrm{T}}, \theta^{\mathrm{T}})^{\mathrm{T}}$ is a $(p+q+S)\times 1$ vector of parameters. This likelihood differs from the one in Section 2 because we need to model $p(\mathbf{d}|\mathbf{c};\gamma)$, the conditional density of the missing covariates given the observed covariates.

Since a joint model for $p(\mathbf{d}|\mathbf{c};\gamma)$ can be difficult to specify in general, we typically model $p(\mathbf{d}|\mathbf{c};\gamma)$ as the product of conditional densities in the same manner as Lipsitz and Ibrahim (1996). That is, we express

$$p(\mathbf{d}|\mathbf{c};\gamma) = p(d_1|\mathbf{c};\gamma_1)\; p(d_2|d_1,\mathbf{c};\gamma_2)\ldots p(d_D|d_1,\ldots,d_{D-1},\mathbf{c};\gamma_D),$$

where $\gamma = (\gamma_1^{\mathrm{T}},\ldots,\gamma_D^{\mathrm{T}})^{\mathrm{T}}$ is the $q\times 1$ vector of parameters for $p(\mathbf{d}|\mathbf{c};\gamma)$. While this provides a simplified model for $p(\mathbf{d}|\mathbf{c};\gamma)$, it does have some drawbacks. First, bias can be introduced in estimation of $\gamma$. Second, the order in which conditional densities are taken can influence the estimation of $\psi$; however, it has been demonstrated that the estimates are quite robust to the order of the conditioning (Lipsitz and Ibrahim, 1996; Ibrahim *et al.*, 1999a,b, 2001).

Since $\mathbf{d}^{\mathrm{miss}}$ and $\mathbf{b}$ are not observed, we base our inference on the observed-data likelihood

$$L(\phi;\mathbf{y},\mathbf{c},\mathbf{d}^{\mathrm{obs}}) = \int \sum_{\mathbf{d}^{\mathrm{miss}}} p(\mathbf{y},\mathbf{d},\mathbf{b}|\mathbf{c};\phi)\,\mathrm{d}\mathbf{b}\;,$$

where the summation is over all possible combinations of the missing discrete variables. The observed-data likelihood in the missing covariate case is even more difficult to work with than that in the complete-covariate case, but the complete-data likelihood remains relatively easy to manipulate. Thus, data augmentation methods, such as the EM algorithm, are the most natural methods to use to obtain estimates of $\phi$. In the following two sections, we outline our approach to estimation of $\phi$ and $\mathrm{Var}(\widehat{\phi})$.

### 3.1 *Estimation of $\phi$*

Treating both $\mathbf{d}^{\mathrm{miss}}$ and $\mathbf{b}$ as missing data, we can find the maximum likelihood estimate of $\phi$ using the EM algorithm (Dempster *et al.*, 1977). However, as in the complete-covariate case, the integral in the E-step is intractable for most generalized linear mixed effects models. To remedy this, we propose using a Laplace EM algorithm (Steele, 1996), handling the missing covariate values with the Method of Weights (Ibrahim, 1990).

Let $\ell(\phi;\mathbf{y},\mathbf{c},\mathbf{d},\mathbf{b}) = \log p(\mathbf{y},\mathbf{d},\mathbf{b}|\mathbf{c};\phi)$, then the E-step at the $(m+1)$th iteration of the EM algorithm calculates

$$\begin{aligned}
\mathsf{D}_\phi Q(\phi|\phi^{(m)}) &= E\{\mathsf{D}_\phi \ell(\phi;\mathbf{y},\mathbf{c},\mathbf{d},\mathbf{b})|\mathbf{y},\mathbf{c},\mathbf{d}^{\mathrm{obs}},\phi^{(m)}\}\\
&= \int \sum_{\mathbf{d}^{\mathrm{miss}}} \mathsf{D}_\phi \ell(\phi;\mathbf{y},\mathbf{c},\mathbf{d},\mathbf{b})\; p(\mathbf{d}^{\mathrm{miss}},\mathbf{b}|\mathbf{y},\mathbf{c},\mathbf{d}^{\mathrm{obs}};\phi^{(m)})\,\mathrm{d}\mathbf{b}\\
&= \int \mathsf{D}_\phi \ell_w(\phi;\tilde{\mathbf{y}},\tilde{\mathbf{c}},\tilde{\mathbf{d}},\mathbf{b})\; p(\mathbf{b}|\mathbf{y},\mathbf{c},\mathbf{d}^{\mathrm{obs}};\phi^{(m)})\,\mathrm{d}\mathbf{b}, \qquad (3.1)
\end{aligned}$$

where $\ell_w(\phi;\tilde{\mathbf{y}},\tilde{\mathbf{c}},\tilde{\mathbf{d}},\mathbf{b}) = E\{\ell(\phi;\mathbf{y},\mathbf{c},\mathbf{d},\mathbf{b})|\mathbf{b},\mathbf{y},\mathbf{c},\mathbf{d}^{\mathrm{obs}},\phi^{(m)}\}$ is a weighted complete-data log likelihood that is calculable in closed form because the missing covariate values are discrete. Ibrahim (1990) derived a similar expression in the context of the generalized linear model. The vectors $\tilde{\mathbf{y}}$, $\tilde{\mathbf{c}}$,and $\tilde{\mathbf{d}}$ are augmented versions of $\mathbf{y}$, $\mathbf{c}$, and $\mathbf{d}$ and are discussed in more detail below.

For the Upper Cape Cod study, there is one discrete covariate ($d_i = s_i$) and three continuous covariates ($\mathbf{c}_i = (a_i, \mathbf{x}_i^{\mathrm{T}})^{\mathrm{T}}$). We model $p(\mathbf{y},\mathbf{b}|\psi)$ with (2.12) and (2.13) and model $p(s_i|\mathbf{c}_i;\gamma)$ as

$$\mathrm{logit}\{P(s_i|\mathbf{c}_i;\gamma)\} = (1\;\mathbf{c}_i^{\mathrm{T}})\gamma.$$

The weighted complete-data log likelihood is given by

$$\ell_w(\phi; \tilde{\mathbf{y}}, \tilde{\mathbf{c}}, \tilde{\mathbf{d}}, \mathbf{b}) = \tilde{\mathbf{y}}^{\mathrm{T}}\mathbf{W}(\tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{Z}}\mathbf{b}) - \mathbf{1}^{\mathrm{T}}\mathbf{W}\,\mathrm{B}(\tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{Z}}\mathbf{b}) + \tilde{\mathbf{s}}^{\mathrm{T}}\mathbf{W}\tilde{\mathbf{X}}_{-s}\boldsymbol{\gamma} - \mathbf{1}^{\mathrm{T}}\mathbf{W}\,\mathrm{B}(\tilde{\mathbf{X}}_{-s}\boldsymbol{\gamma}) - \tfrac{1}{2}\mathbf{b}^{\mathrm{T}}\mathbf{D}_{\boldsymbol{\theta}}^{-1}\mathbf{b} - \tfrac{1}{2}\log|\mathbf{D}_{\boldsymbol{\theta}}|\,, \tag{3.2}$$

where $\tilde{\mathbf{X}}_{-s}$ is the matrix $\tilde{\mathbf{X}}$ with the column for smoking status removed and $\mathbf{b} = (\mathbf{b}_a^{\mathrm{T}}, \mathbf{b}_{\mathbf{x}}^{\mathrm{T}})^{\mathrm{T}}$ is partitioned to conform to the partitioning of $\mathbf{Z}$.

In general, the $\tilde{n} \times p$ matrix $\tilde{\mathbf{X}}$ has rows $\tilde{\mathbf{x}}_{i(j)}^{\mathrm{T}} = [1,\ \tilde{\mathbf{d}}_{i(j)}^{\mathrm{T}},\ \mathbf{s}_i^{\mathrm{T}}]$, where $\tilde{n} = \sum_i n_i$ and $\tilde{\mathbf{d}}_{i(j)}$ is the vector $\mathbf{d}_i$ with $\mathbf{d}_i^{\mathrm{miss}} = \mathbf{d}_{i(j)}^{\mathrm{miss}}$, the $j$th possible combination of values for $\mathbf{d}_i^{\mathrm{miss}}$. Thus, $\tilde{\mathbf{X}}$ is constructed by replacing the $i$th original row of $\mathbf{X}$ by $n_i$ rows, with each new row containing one of the $n_i$ possible combinations for the $\mathbf{d}_i^{\mathrm{miss}}$ values. The values $\mathbf{d}_i^{\mathrm{obs}}$ remain unchanged over the $n_i$ observations. The matrix $\mathbf{W}$ is a $\tilde{n} \times \tilde{n}$ diagonal matrix with elements

$$w_{i(j)} = \begin{cases} \Pr(\mathbf{d}_i^{\mathrm{miss}} = \mathbf{d}_{i(j)}^{\mathrm{miss}} | \mathbf{b}, \mathbf{y}, \mathbf{c}, \mathbf{d}^{\mathrm{obs}}; \phi^{(m)}) & \text{if observation } i \text{ has missing values} \\ 1 & \text{otherwise} \end{cases} \tag{3.3}$$

where

$$\Pr(\mathbf{d}_i^{\mathrm{miss}} = \mathbf{d}_{i(j)}^{\mathrm{miss}} | \mathbf{b}, \mathbf{y}, \mathbf{c}, \mathbf{d}^{\mathrm{obs}}; \phi^{(m)}) = \frac{p(y_i, \tilde{\mathbf{d}}_{i(j)}, \mathbf{b} | \mathbf{c}_i; \phi^{(m)})}{\sum_{k=1}^{n_i} p(y_i, \tilde{\mathbf{d}}_{i(k)}, \mathbf{b} | \mathbf{c}_i; \phi^{(m)})} \tag{3.4}$$

and the summation is over the $n_i$ possible values of $\mathbf{d}_i^{\mathrm{miss}}$. Equation (3.4) follows from a direct application of Bayes' Theorem. Since they contain completely observed data, the $\tilde{n} \times M$ matrix $\tilde{\mathbf{Z}}$ and $\tilde{n} \times 1$ vector $\tilde{\mathbf{y}}$ are created by replacing the $i$th row of $\mathbf{Z}$ and $\mathbf{y}$ with $n_i$ copies of $\mathbf{z}_i^{\mathrm{T}}$ and $y_i$, respectively.

Since (3.1) cannot be calculated in closed form, we follow Steele (1996) and use an approximation based on a fully exponential Laplace approximation (Tierney *et al.*, 1989). The approximation yields

$$\widehat{\mathsf{D}}_{\phi} Q(\phi | \phi^{(m)}) = \left\{ \mathsf{D}_{\phi} \ell_w(\phi; \tilde{\mathbf{y}}, \tilde{\mathbf{c}}, \tilde{\mathbf{d}}, \mathbf{b}) + C(\phi; \tilde{\mathbf{y}}, \tilde{\mathbf{c}}, \tilde{\mathbf{d}}, \mathbf{b}) \right\} \Big|_{\mathbf{b} = \widehat{\mathbf{b}}(\phi^{(m)})}\,, \tag{3.5}$$

where

$$\widehat{\mathbf{b}}(\phi) \equiv \mathrm{argmax}_{\mathbf{b}}\, p(\mathbf{y}, \mathbf{c}, \mathbf{d}^{\mathrm{obs}}, \mathbf{b}; \phi).$$

The correction term $C(\phi; \tilde{\mathbf{y}}, \tilde{\mathbf{c}}, \tilde{\mathbf{d}}, \mathbf{b})$ is similar to (2.19) with appropriate changes. For example, $\mathbf{A}$ is replaced by $\mathbf{A}_w = -\mathsf{H}_{\mathbf{b}} \ell_w(\phi; \tilde{\mathbf{y}}, \tilde{\mathbf{c}}, \tilde{\mathbf{d}}, \mathbf{b})$. In calculating the correction terms, we replace the EM weights, $w_{i(j)}$ with a first-order Taylor series expansion about $\widehat{\mathbf{b}}(\phi^{(m)})$. A detailed derivation of (3.5) and, in particular, $C(\phi; \tilde{\mathbf{y}}, \tilde{\mathbf{c}}, \tilde{\mathbf{d}}, \mathbf{b})$ is given in an appendix found on Biostatistics Online.

It is well-known that the EM algorithm does not automatically provide a method for estimating $\mathrm{Var}(\widehat{\phi})$. Because it relies on the same code used to run the EM algorithm and has been shown to be at least as accurate as the SEM method (Meng and Rubin, 1991), we use the forward-difference method of Jamshidian and Jennrich (2000).

### 3.2 *Prediction of* $\mathbf{b}$

To estimate the surface $S(\mathbf{x})$, we need to predict the random effects $\mathbf{b}$. It is common to predict $\mathbf{b}$ with $\mathbf{b}^* = E(\mathbf{b}|\mathbf{y}, \mathbf{c}, \mathbf{d}^{\mathrm{obs}}, \widehat{\phi})$, the conditional expectation of $\mathbf{b}$ given the observed data evaluated at the maximum likelihood estimate for $\phi$. Since, for the generalized linear mixed effects model, this expectation typically cannot be calculated in closed form, we approximate it with a Laplace approximation. The spirit of the approximation is similar to that used to approximate the E-step of the EM algorithm described in Section 3.1. Details are provided in an appendix found on Biostatistics Online.

## 4. Analysis of the upper Cape Cod cancer data

In this section we return to the Upper Cape Cod prostate cancer data collected by the MDPH between 1987 and 1994. The response variable is the presence or absence of prostate cancer ($y_i = 1$ and $y_i = 0$, respectively) among $n = 3191$ male cancer patients living on the upper portion of Cape Cod. In addition to location of residence, covariate data was collected on age and smoking status. Age is categorized into age decade $(1, 2, \ldots, 10)$, and a subject's smoking status is recorded as either 'ever smoker' ($s_i = 1$) or 'never smoker' ($s_i = 0$). A plot of the geo-coded location data is shown in Figure 1. Prostate cancer cases are plotted using solid dots, and other cancer cases are plotted using empty dots. To preserve confidentiality the points have been jittered. A summary of the other variables is presented in Table 1. While age and location are completely observed, smoking status was not obtained for approximately 29% of the subjects.

The primary goal of this analysis is to assess the geographic distribution of prostate cancer risk among the population of Upper Cape Cod, adjusting for the effects of smoking and age. To estimate this risk surface, we need a comparison group of control subjects. This group should be a sample from the population of people living on Upper Cape Cod and at risk for being diagnosed with prostate cancer between 1987 and 1994. Since a random sample from this population is not available, we use as control subjects the male patients diagnosed with other cancers during the same time period. In using this set of controls, we must assume that the risk of other cancers is not associated with the location of the patient's residence.

The issues surrounding relative cancer mapping, and proportional incidence studies in general, are important ones. Breslow and Day (1987, p. 45) note that proportional incidence studies can be useful in the initial stages of an investigation. However, using other cancer types as controls can be problematic. In particular, interpretation of the results can be difficult since an increased proportional risk can be the result of either an increased absolute risk of prostate cancer or a decreased absolute risk for other types of cancer. However, if we are comfortable with the assumption that incidence of other types of cancer is not related to geography, then the data can be viewed as if they arose from a case-control study in which the other cancer cases are assumed to represent an unbiased sample of the population at risk (Breslow and Day, 1987, p. 115).

In general, we should exclude from any analysis those cancer types that are known to be related to geography (Breslow and Day, 1987, p. 115) and those cancers that are similar in etiology to the cancer under study. For example, Best and Wakefield (1999) exclude cancers of the uterus and ovary in a proportional incidence model for breast cancer. In our analysis of the Cape Cod data, we have included all other cancer types in the control group. No cancer types were excluded because we had no evidence that other cancers are clearly associated with geography and no other cancers that were recorded in the cancer registry had an etiology similar to that of prostate cancer.

We model these data with the following additive logistic regression model

$$\text{logit}\{P(y_i = 1)\} = \beta_0 + \beta_s s_i + f(a_i) + S(\mathbf{x}_i) \tag{4.1}$$

where $\mathbf{x}_i = (\text{longitude}_i, \text{latitude}_i)$, $f$ is a smooth univariate function of age and $S$ is a smooth bivariate function of longitude and latitude. We estimate (4.1) using the logistic mixed model defined by (2.7)–(2.10). Kelsall and Diggle (1998) and Diggle *et al.* (1998) have fit similar models to case-control data.

For the age variable, we use $K_a = 7$ knots placed at the each of the interior points of the range of age, $\{2, 3, 4, 5, 6, 7, 8\}$. For the geographical data, we use $K_{\mathbf{x}} = 33$ knots chosen using a space-filling algorithm from the `S-PLUS` module `FUNFITS` for fitting thin plate splines. The random effects, $\mathbf{b} = (\mathbf{b}_a^{\mathrm{T}}, \mathbf{b}_{\mathbf{x}}^{\mathrm{T}})^{\mathrm{T}}$, are assumed to follow a Gaussian distribution, $\mathbf{b} \sim N(\mathbf{0}, \mathbf{D}_{\boldsymbol{\theta}})$, where $\mathbf{D}_{\boldsymbol{\theta}} = \text{diag}(\sigma_a^2 \mathbf{1}_{K_a}, \sigma_{\mathbf{x}}^2 \mathbf{1}_{K_{\mathbf{x}}})$. To improve the likelihood surface, we reparameterize the model using $\boldsymbol{\theta} = (\theta_a, \theta_{\mathbf{x}})^{\mathrm{T}} = (\log \sigma_a^2, \log \sigma_{\mathbf{x}}^2)^{\mathrm{T}}$. To fit this model using the algorithm presented in Section 3, we need to additionally specify a model for the conditional probability of being a smoker given the completely

Table 2. *Complete-case and Method of Weights analyses of the Upper Cape Cod cancer data*

| | Method of Weights | | Complete case | |
|---|---|---|---|---|
| Parameter | Estimate | (s.e.) | Estimate | (s.e.) |
| Intercept | −0.35 | (0.118) | −0.61 | (0.120) |
| smoke | −0.90 | (0.122) | −0.89 | (0.106) |
| $\theta_{\text{age}}$ | 5.20 | | 5.01 | (0.463) |
| $\theta_{\text{geog}}$ | 3.77 | | 3.57 | (0.367) |

observed covariates, age and residence. We model smoking status using the following logistic regression model:

$$\text{logit}\{P(s_i = 1)|a_i, \mathbf{x}_i\} = \gamma_0 + \gamma_1 a_i + \gamma_2 \text{longitude}_i + \gamma_3 \text{latitude}_i.$$

To run the complete-case analysis, we fit model (4.1), excluding the 940 subjects with missing smoking data. The resulting generalized linear mixed model was estimated using the `GLIMMIX` macro in `SAS`. To run the Method of Weights analysis, we fit model (4.1) using the observed data from all 3191 subjects. The model was estimated using programs written in `C` and run on a Sun workstation.

The upper-right panel of Figure 1 plots the estimated odds ratio surfaces for geography from the complete-case analysis. The odds ratio for the geographic component is plotted relative to the geometric mean odds. There appears to be an area of increased risk to the southeast of the MMR, between the towns of Falmouth and Hyannis. The lower-left panel of Figure 1 is the corresponding plot based on estimates from the Method of Weights analysis. The estimated geographic component is strikingly different. In particular, the area of elevated risk has extended to the north, but the maximum risk has decreased. While similar regions are elevated in both analyses, this shift may be important from the perspective of an hypothesis-generating analysis.

To assess the sensitivity of the results to the choice of $\rho$, we separately fit models using values of $\rho$ ranging from $\hat{\rho} = \max_{1 \leqslant i,j \leqslant n} \|\mathbf{x}_i - \mathbf{x}_j\|$ to $\hat{\rho} = \frac{1}{20} \max_{1 \leqslant i,j \leqslant n} \|\mathbf{x}_i - \mathbf{x}_j\|$. The smoothness of the estimated surfaces did depend on the choice of $\rho$, with larger values of $\hat{\rho}$ yielding smoother surfaces. However, the scale of the maps was virtually the same for all choices of $\hat{\rho}$.

The lower-right panel of Figure 1 plots the estimated odds ratio for age relative to the mean age in the cohort, approximately 77 years. The solid line corresponds to the Method of Weights estimate, while the dashed line corresponds to the complete-case EM estimate. The primary difference between the two estimates occurs in the older age groups. The complete-case estimate shows a risk that continues to increase after but the rate slows after age 75. The Method of Weights estimate shows that the risk begins to decrease risk around an age of 80 years. Table 2 shows the parameter estimates for smoking from both the complete-case EM algorithm and the Method of Weights algorithm. We see that prostate cancer patients are less likely to be smokers than patients with other cancers. This is likely driven by the relatively large number of lung cancer patients in the control group.

We note that this analysis provides only a first attempt at examining these data. Due to the long latency period of most cancers and the difficulties in interpreting results of proportional incidence studies, this analysis cannot provide conclusive evidence of an association between location of residence and onset of prostate cancer. However, we believe that it can be used to generate further hypotheses about potential environmental effects. We also note that we chose to fit models for prostate cancer risk after seeing elevated rates in the MDPH report (Massachusetts Department of Public Health, 1997). While formal hypothesis testing may be compromised as a result, estimation of the risk surface still provides a more refined estimate of risk than do methods based on town- or census tract-level prostate cancer rates.

## 5. Discussion

In this paper we have presented a method for fitting covariate-adjusted cancer incidence maps in the presence of missing covariate data. By fitting these maps using a generalized linear mixed effects model, we are able to use the EM algorithm to obtain estimates of both the fixed effects and variance components which control the amount of smoothing. Thus, we have a method that can both handle missing data and provide an automatic method for choosing the 'optimal' amount of smoothing.

Important similarities and differences exist between the method presented here and the complete-case alternative. Both approaches treat the additive model (2.2) as a mixed effects model. While either standard software (e.g. PQL via the `GLIMMIX` macro in `SAS`) or custom software (e.g. the EM algorithm via C or FORTRAN programs) can be used to fit the model to the complete-cases, custom programs must be written to use the Method of Weights algorithm described here. The potential benefits of using our approach lie in the validity of the estimates. The complete-case analysis can yield biased estimates if the missingness mechanism depends on the response variable (Jones, 1996). The Method of Weights approach, however, will yield unbiased estimates in this situation. Conversely, the complete-case analysis can provide unbiased estimates when the missingness mechanism depends on the missing covariate values; a property that is not shared by our method.

Several additional points are worth discussing. The first regards the Laplace approximation to the intractable integral in the E-step of the EM algorithm. The dimension of the integral is the same as the total number of knots for estimating the smooth curves. Thus, this integral may be of quite high dimension, even when using a reduced set of knots as we do. The question arises, then, as to the validity of the approximation. Clearly, in such cases we need to have a sufficiently large sample size in order for the approximation to be valid. However, this is exactly the situation in which Monte Carlo methods are tedious to implement and when we would advocate using this method. Though not implemented here, it would be worthwhile to assess how the methodology works as the number of knots are varied. Increasing the number of knots may allow for a smoother estimated curve or surface, but may also yield a poorer Laplace approximation. An alternative approach is that taken in French (2000) where a Monte Carlo EM algorithm (Wei and Tanner, 1990) is implemented. While such an approach may be able to provide a more accurate approximation, sampling from the distribution of the random effects given the observed data can be slow even when the data set is relatively small. One benefit of the method proposed here is that we can easily handle large data sets such as the Cape Cod cancer data.

A second point concerns estimation of standard errors for the fitted curves and surfaces. While we use the FDM method (Jamshidian and Jennrich, 2000) to obtain estimated standard errors for the fixed effects, we do not provide an estimate of the standard error for $\hat{f}(a)$ and $\hat{S}(\mathbf{x})$. We are currently studying the use of bootstrap and MCMC methods for standard error estimation.

Finally, we would like to obtain a method for testing the hypothesis that the surface is planar. Since a planar surface corresponds to $\sigma^2 = 0$, such a test can be obtained by testing $H_0 : \sigma^2 = 0$. Crainiceanu *et al.* (2004) use this type of approach in the linear model. Application of these methods to generalized linear mixed effects models is a subject of current research.

## Acknowledgements

The authors would like to thank the Massachusetts Department of Public Health for making the Upper Cape Cod cancer incidence data available for use. The authors would also like to thank Joe Ibrahim and Louise Ryan for providing comments on an earlier draft. This work was supported in part by NIH grant T32 ES07142-18 and was completed while J. L. French was a doctoral student at the Harvard School of Public Health.

## REFERENCES


BECK, N. AND JACKMAN, S. (1998). Beyond linearity by default: generalized additive models. *American Journal of Political Science* **42**, 596–627.

BEST, N. AND WAKEFIELD, J. (1999). Accounting for inaccuracies in population counts and case registration in cancer mapping. *Journal of the Royal Statistical Society, Series A* **162**, 363–382.

BOOTH, J. G. AND HOBERT, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B* **61**, 265–285.

BRESLOW, N. E. AND CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

BRESLOW, N. E. AND DAY, N. E. (1987). *Statistical Methods in Cancer Research*. Lyon: International Agency for Research on Cancer.

BRUMBACK, B. A. AND RICE, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* **93**, 961–1006.

BRUMBACK, B. A., RUPPERT, D. AND WAND, M. P. (1999). Comment on Shively, Kohn, and Wood. *Journal of the American Statistical Association* **94**, 794–797.

CRAINICEANU, C., RUPPERT, D., CLAESKENS, G. AND WAND, M. P. (2004). Likelihood ratio tests of polynomial regression against a general alternative. *Biometrika*, to appear.

CRESSIE, N. (1993). *Statistics for Spatial Data*. New York: Wiley.

DAVIS, J. M. AND SPECKMAN, P. (1999). A model for predicting maximum and 8 h average ozone in Houston. *Atmospheric Environment* **33**, 2487–2500.

DEMPSTER, A. P., LAIRD, N. M. AND RUBIN, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

DIGGLE, P. J., TAWN, J. A. AND MOYEED, R. A. (1998). Model-based geostatistics (Disc: p326-350). *Applied Statistics* **47**, 299–326.

ENGELS, E. A., ROSENBERG, P. S. AND BIGGAR, R. J. (1999). Zoster incidence in human immunodeficiency virus-infected hemophiliacs and homosexual men, 1984–1997. *Journal of Infectious Diseases* **180**, 1784–1789.

FRENCH, J. L. (2000). Analysis of environmental health data with missing values, Unpublished Sc.D. Thesis.

HASTIE, T. J. AND TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. London: Chapman & Hall.

IBRAHIM, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association* **85**, 765–769.

IBRAHIM, J. G., CHEN, M.-H. AND LIPSITZ, S. R. (1999a). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics* **55**, 591–596.

IBRAHIM, J. G., CHEN, M.-H. AND LIPSITZ, S. R. (2001). Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. *Biometrika* **88**, 551–564.

IBRAHIM, J. G., LIPSITZ, S. R. AND CHEN, M.-H. (1999b). Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *Journal of the Royal Statistical Society, Series B* **61**, 173–190.

JAMSHIDIAN, M. AND JENNRICH, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society, Series B* **62**, 257–270.

JOHNSON, M. E., MOORE, L. M. AND YLVISAKER, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference* **26**, 131–148.

JONES, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association* **91**, 222–230.

KAMMANN, E. E. AND WAND, M. P. (2003). Geoadditive models. *Applied Statistics* **52**, 1–18.

KELSALL, J. E. AND DIGGLE, P. J. (1998). Spatial variation in risk of disease: a nonparametric binary regression approach. *Applied Statistics* **47**, 559–573.

LIN, X. AND ZHANG, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B* **61**, 381–400.

LINTON, O. B. AND HÄRDLE, W. (1996). Estimation of additive regression models with known links. *Biometrika* **83**, 529–540.

LIPSITZ, S. R. AND IBRAHIM, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika* **83**, 916–922.

LITTLE, R. J. A AND RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44**, 226–233.

MASSACHUSETTS DEPARTMENT OF PUBLIC HEALTH (1997). *Cancer Incidence in Massachusetts 1987–1994: City/Town Supplement*. Boston, MA: MDPH.

MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**, 162–170.

MENG, X. L. AND RUBIN, D. B. (1991). Using the EM algorithm to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association* **86**, 899–909.

NYCHKA, D. AND SALTZMAN, N. (1998). Design of air quality monitoring networks. In D., Nychka, Cox, L. and Piegorschs, W. (eds), *Case Studies in Environmental Statistics*, New York: Springer.

NYCHKA, D., HAALAND, P., O'CONNELL, M. AND ELLNER, S. (1998). FUNFITS, data analysis and statistical tools for estimating functions. In Nychka, D., Piegorsch, W. W. and Cox, L. H. (eds), *Case Studies in Environmental Statistics*, New York: Springer.

ROLAND, J., KEYGHOBADI, N. AND FOWNES, S. (2000). Alpine Parnassius butterfly dispersal: effects of landscape and population size. *Ecology* **81**, 1642–1653.

SCHWARTZ, J. (1997). Air pollution and hospital admissions for cardiovascular disease in Tuscon. *Epidemiology* **8**, 371–377.

STEELE, B. M. (1996). A modified EM algorithm for estimation in generalized mixed models. *Biometrics* **52**, 1295–1310.

STEIN, M. L. (1999). *Statistical Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer.

TIERNEY, L., KASS, R. E. AND KADANE, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association* **84**, 710–716.

VERBYLA, A., CULLIS, B. R., KENWARD, M. G. AND WELHAM, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *Applied Statistics* **48**, 269–300.

WAHBA, G. (1990). *Spline Models for Observational Data*. Philadelphia, PA: SIAM.

WALKER, S. (1996). An EM algorithm for nonlinear random effects models. *Biometrics* **52**, 934–944.

WANG, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association* **93**, 341–348.

WEI, G. C. G. AND TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85**, 699–704.

[*Received June 13, 2001; revised May 31, 2002; accepted for publication June 11, 2002*]