Geoadditive models

BY E. E. KAMMANN AND M. P. WAND

*Department of Biostatistics, School of Public Health, Harvard University, 665
Huntington Avenue, Boston, Massachusetts 02115, U.S.A.*

31st July, 2000

SUMMARY

A study into geographical variability of reproductive health outcomes (e.g. birthweight) in Upper Cape Cod, Massachusetts, USA, benefits from geostatistical mapping or *kriging*. However, also observed are a number of continuous covariates (e.g. maternal age) that exhibit pronounced non-linear relationships with the response variable. To properly account for such effects we merge kriging with additive models to obtain what we call *geoadditive models*. The mergence becomes effortless by expressing both as linear mixed models. The resulting mixed model representation for the geoadditive model allows for fitting and diagnosis using standard methodology and software.

Keywords: Additive models; Disease Mapping; Geostatistics; Kriging; Mixed Models; Nonparametric Regression; Penalised Splines; Restricted Maximum Likelihood.

1 Introduction

Geostatistics is concerned with the problem of producing a map of a quantity of interest over a particular geographical region based on, usually noisy, measurements taken at a set of locations in the region. Figure 1 provides an illustration. The left-hand panel shows residuals from a fitted regression model in which birthweight was regressed against several infant and maternal attributes from an environmental health study in Upper Cape Cod, Massachusetts, USA (see Section 2). The right hand panel is a "map" of the residuals obtained via the geostatistical method known as *kriging*. It provides an informative summary of the geographical variation in mean birthweight over the region and, in particular, shows possible 'hot spots' of adverse health outcomes.

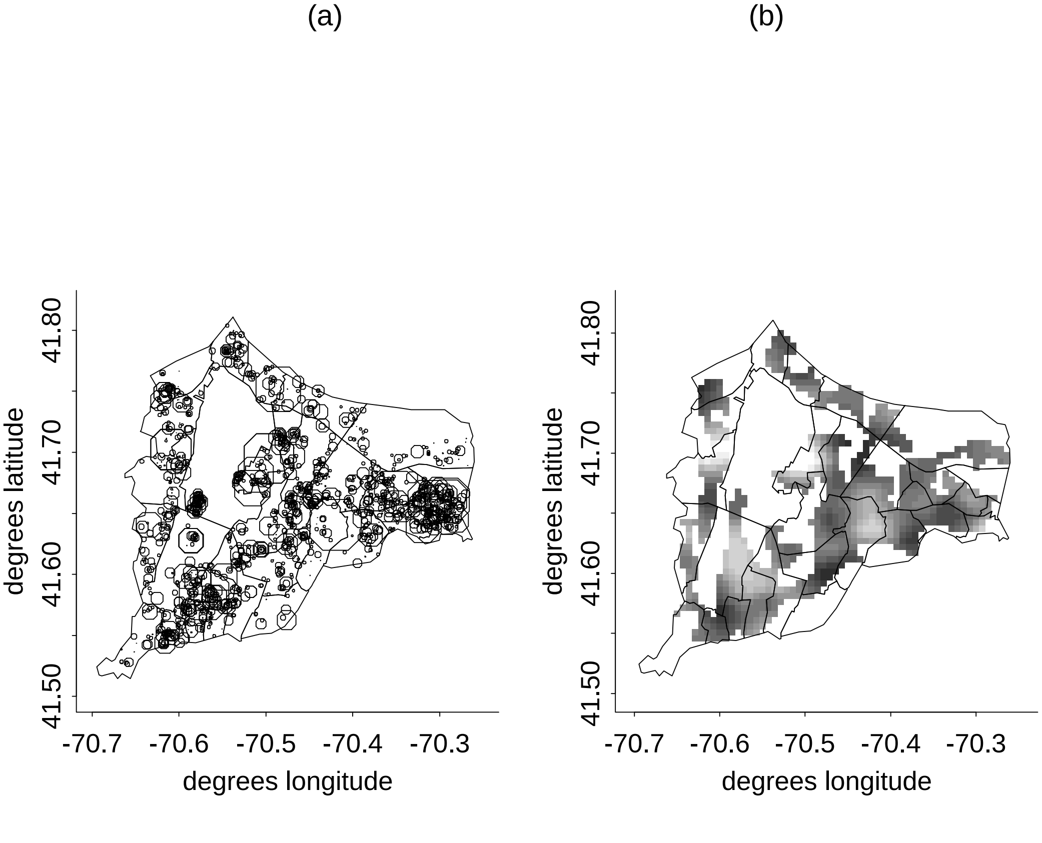


Figure 1: (a) Residuals from an additive model fit of birthweight to several covariates plotted geographically. The size of the circle indicates the size of the residual. (b) A map of the data in (a) obtained using kriging.

The data used in Figure 1 (a) were obtained as part of a study into geographical variation in health outcomes in Upper Cape Cod. Details of the data are given in Section 2. Investigations of this nature are very common and a recent article in *The New Yorker* magazine (Gawande, 1999) reported that, in 1998, the state of Massachusetts responded to more than three thousand disease cluster alarms, most of which concerned cancer. The article was also quite critical of such investigations, pointing out that not one cancer cluster has been convincingly identified. One of the main reasons for this is the lengthy duration of time before the onset of clinical symptoms of many types of cancer. The Upper Cape Cod investigation began as cancer cluster studies, but more recently has turned to reproductive outcomes such as birthweight. Reproductive outcomes have the advantage of being sensitive to recent exposures.

Even for perfect measures of adverse health, kriging alone will not properly address the question of environmental causality. For example, a region with lower income levels is also more likely to have higher levels of adverse health outcomes. The Upper Cape Cod study aims to redress this problem by obtaining data on all other available attributes and accounting for them in the mapping. Figure 1 represents a cursory attempt to control for covariates. As mentioned above, a regression model was fit to the attributes and then residuals were mapped. But, ideally, these processes would be done simultaneously. The extension of kriging, sometimes known as *universal* kriging (e.g. Cressie, 1993; Hobert, Altman and Schofield, 1997), allows for the incorporation of covariates. However, linearity of the covariate effects is usually assumed. This is not satisfactory for the motivating example since, for instance, maternal age has a non-linear effect on gestational age. Indeed, the regression model used to produce Figure 1 is an *additive* model (e.g. Hastie and Tibshirani, 1990) which permits general smooth functional covariate effects. Our goal is therefore to simultaneously map reproductive outcomes such as birthweight and gestational age while accounting for non-linear covariate effects under the assumption of additivity. The resulting models represent a fusion of geostatistical and additive models, hence the name *geoadditive models*.

There are several ways to combine the ideas of geostatistics and additive modelling. Our research has lead to models that have the following advantages:

(1) seamless; due to using a mixed model representation of both kriging and additive models,

(2) model-based and likelihood-driven; our geoadditive model is simply a linear mixed model and, under Gaussian distributional assumptions, lends itself to estimation of all parameters using (restricted) maximum likelihood and testing via the likelihood ratio paradigm,

(3) low-rank, as defined by Hastie (1996); meaning that the number of basis functions used to construct the function estimates does not grow with the sample size; which is vitally important for disease mapping applications, including the motivating problem, where the data often number in the thousands; and

(4) implementable using standard software. With some simplification in the kriging component we are able to express the model as a sub-class of mixed models commonly known as *variance component models* (e.g. Searle, Casella and McCulloch, 1992). This leads to enormous reduction in computational complexity and allows for the direct use of standard software such as `PROC MIXED` in `SAS` and `lme()` in `S-PLUS`.

Worthwhile background reading for this paper is a recent article on model-based geostatistics by Diggle, Tawn and Moyeed (1998) where pure kriging (i.e. no covariates) is the focus. Our paper inherits some of its aspects: model-based and with mixed model connections. In particular the comment by Bowman (1998) in the ensuing discussion suggested that additive modelling would be a worthwhile extension. This paper essentially follows this suggestion. However, this paper is not the first to combine the notions of geostatistics and additive modelling. References known to us are Kelsall and Diggle (1998), Durbán Reguera (1998) and Durbán, Hackett, Currie and Newton (2000). Nevertheless, we believe that our approach has a number of attractive features (see (1)-(4) above), not all shared by these references.

Section 2 describes the motivating application and data in detail. Section 3 shows how one can express additive models as a mixed model, while Section 4 does the same for kriging and merges the two into the geoadditive model. Issues concerning the amount of smoothing are discussed in Section 5 and inferential aspects are treated in Section 6. Our analysis of the Upper Cape Cod reproductive data is presented in Section 7. Section 8 discusses extension to the generalised context.We close the paper with some disussion in Section 9.

2 Description of the application and data

A number of environmental health studies have taken place in the region of Massachusetts known as Upper Cape Cod since elevated cancer rates were observed there in the mid-1980s. Several possible sources have been identified and include fuel dumping at a large military reservation, pesticide use in cranberry bogs and poly-chlorinated biphenyl in water pipes. However, the studies have been largely inconclusive.

In the late 1990s the Department of Public Health, Commonwealth of Massachusetts, commissioned a new study into geographical variation of health outcomes in Upper Cape Cod. In the latest phase reproductive outcomes, birthweight and gestational age, have been considered. Birthweight is measured on nearly all newborns, and is sensitive to recent exposures, thus facilitating the determination of exposures of biological importance. For example, a 170-200 gramme decrease in mean birthweight may be seen in babies whose mothers smoke over 16 cigarettes per day during pregnancy compared with those who do not smoke. Similar arguments can be made for studying gestational age.

From a statistical viewpoint, birthweight and gestational age have the advantage of being continuous. Figure 2 gives histograms for these variables corresponding to the Upper Cape Cod data set described below. Apart from the relatively small number of light or premature births both variables are free of any significant skewness. This leads to a simpler model and analysis since the Gaussian assumption is more tenable.

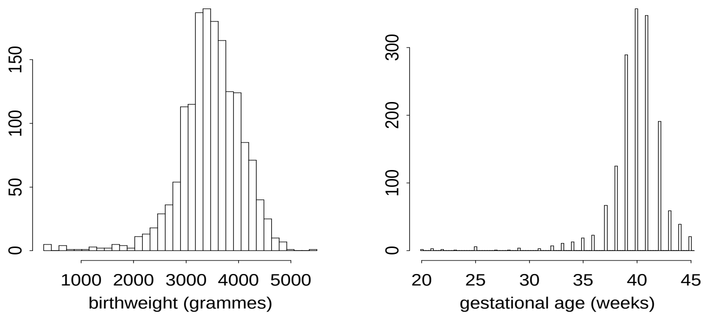


Figure 2: Histograms of birthweight and gestational age for the Upper Cape Cod reproductive data described in this section.

The Upper Cape Cod reproductive data correspond to all 1630 births in 1990 across five towns; Barnstable, Bourne, Falmouth, Mashpee and Sandwich. Apart from geographical location (longitude and latitude) and the outcome variables birthweight (grammes) and gestational age (weeks), there are 39 covariates. A preliminary analysis showed that many have no significant association with birthweight or gestational age. Those that are significantly associated with either outcome include maternal age, years of education, number of cigarettes per day and number of drinks per week. Table 1 lists all other variables that exhibited some association with the outcome variables, together with abbreviated names that are used in the analysis summaries in Section 7.

3 Penalised spline additive models

The first half of our model formulation involves a low-rank mixed model representation of additive models (e.g. Brumback *et al.*, 1999). For simplicity we will describe the case of two additive components first. Suppose that (s_i, t_i, y_i), $1 \leqslant i \leqslant n$, represents measurements on two predictors s and t and a response variable y. The additive model for these data is

$$\boxed{y_i = \beta_0 + f(s_i) + g(t_i) + \varepsilon_i} \tag{1}$$

abbreviation	description
infant covariates	
male	indicator for infant being male
black	indicator for infant being black
asian	indicator for infant being Asian
plurality	1=single, 2=twin etc
maternal covariates	
parity	number of live births from mother
diabetes	indicator for diabetes
prenatal visits	number of prenatal care visits
preg. hyperten.	pregnancy-related hypertension
incomp. cervix	indicator for incomplete cervix
eclampsia	indicator for eclampsia
light prev. birth	previous pre-term infant
heavy prev. birth	previous infant ⩾4000 grammes
psychiatric	indicator for psychiatric disorder
renal disease	indicator for renal disease
uterine bleeding	indicator for uterine bleeding

Table 1: Covariates that had some association with birthweight and/or with gestational age according to a preliminary analysis. The abbreviated names are used in the analysis summaries in Section 7.

where $\underline{f}$ and $\underline{g}$ are smooth, but otherwise unspecified, functions of s and t respectively. A penalised spline version of (1) involves fitting

$$y_i = \beta_0 + \boxed{\beta_s s_i + \sum_{k=1}^{K_s} b_k^s \underline{(s_i - \kappa_k^s)}_+} + \boxed{\beta_t t_i + \sum_{k=1}^{K_t} b_k^t \underline{(t_i - \kappa_k^t)_+}} + \varepsilon_i \qquad (2)$$

via least squares, but with penalisation of the knot coefficients b_k^s and b_k^t (e.g. Marx and Eilers, 1998; Ruppert and Carroll, 2000). Here $\kappa_1^s, \ldots, \kappa_{K_s}^s$ and $\kappa_1^t, \ldots, \kappa_{K_t}^t$ are knots in the s and t directions respectively. Rules such as one knot for every 3-4 unique predictor values, up to a maximum of 20-40 knots, are commonly used; although the sensitivity to this choice is quite low (Ruppert, 2000). A key connection is that penalisation of the b_k^s and b_k^t is equivalent to treating them as random effects in a mixed model. Specifically, if we define $\underline{\boldsymbol{\beta} = [\alpha, \beta_s, \beta_t]}^\mathsf{T}$, $\underline{\mathbf{b}} = \underline{[b_1^s, \ldots, b_{K_s}^s, b_1^t, \ldots, b_{K_t}^t]}^\mathsf{T}$,

$$\underline{\mathbf{X} = [1\ s_i\ t_i]}_{1 \leqslant i \leqslant n}, \quad \underline{\mathbf{Z} = [\mathbf{Z}_s | \mathbf{Z}_t]}$$

$$\mathbf{Z}_s = [(s_i - \kappa_k^s)_+]_{1 \leqslant i \leqslant n, 1 \leqslant k \leqslant K_s} \quad \text{and} \quad \mathbf{Z}_t = [(t_i - \kappa_k^t)_+]_{1 \leqslant i \leqslant n, 1 \leqslant k \leqslant K_t} \qquad (3)$$

then penalised least squares is equivalent to best linear unbiased prediction in the mixed model:

$$\boxed{\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon},} \quad E\begin{bmatrix}\mathbf{b} \\ \boldsymbol{\varepsilon}\end{bmatrix} = \mathbf{0}, \quad \text{Cov}\begin{bmatrix}\mathbf{b} \\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\sigma_s^2\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_t^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_\varepsilon^2\mathbf{I}\end{bmatrix}. \qquad (4)$$

Note that (4) is a variance components model since the covariance matrix of $\left[\mathbf{b}^{\mathsf{T}}\ \boldsymbol{\varepsilon}^{\mathsf{T}}\right]^{\mathsf{T}}$ is diagonal. This is one of the simplest mixed model structures and can be readily fitted using standard software.

The variance ratio $\sigma_\varepsilon^2/\sigma_s^2$ acts as a smoothing parameter in the s direction. Intuitively, a very small value of σ_s^2 leads to overfitting of the truncated lines $(s-\kappa_k)_+$ while a very large value leads to a linear fit. Similar comments apply to the t direction. Smoother fits can be obtained via higher degree spline bases. Alternative representations in terms of B-spline and Demmler-Reinsch bases also exist (Eilers and Marx, 1996; Nychka and Cummins, 1996).

Penalised spline additive models are based on *low-rank* smoothers, as defined by Hastie (1996). A precise mathematical definition can be given in terms of the rank of the 'hat' or 'smoother' matrices, but essentially it corresponds to the number of basis functions staying fixed at $K_s + K_t + 3$, usually about 40–60, regardless of the sample size. For very large n this leads to a computationally less intensive fit with little degradation in the estimator (Hastie, 1996).

The extension to higher numbers of additive components is straightforward. Linear terms are easily incorporated into the model through the $\mathbf{X}\boldsymbol{\beta}$ component. As we will show in subsequent sections, this mixed model representation has several benefits in terms of model formulation, fitting and diagnosis.

4 Geostatistical extension

Incorporation of a geographical component can be achieved by expressing kriging as a linear mixed model and merging it with an additive model such as (4) to obtain a single mixed model, which we call the *geoadditive model*.

Suppose that the data are $(\mathbf{x}_i, y_i)$, $1 \leqslant i \leqslant n$, where the y_i's are scalar and $\mathbf{x}_i \in \mathbb{R}^2$ represents geographical location. The simple universal kriging model for such data is

$$y_i = \beta_0 + \boldsymbol{\beta}_1^{\mathsf{T}}\mathbf{x}_i + S(\mathbf{x}_i) + \varepsilon_i \tag{5}$$

where $\{S(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^2\}$ is a stationary zero-mean stochastic process and the ε_i are assumed to be independent zero mean random variables with common variance σ_ε^2 and distributed independently of S (e.g. Cressie, 1993). Prediction at an arbitrary location $\mathbf{x}_0 \in \mathbb{R}^2$ is typically done through an expression of the form

$$\widehat{y}(\mathbf{x}_0) = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1^{\mathsf{T}}\mathbf{x}_0 + \widehat{S}(\mathbf{x}_0)$$

where $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\beta}}_1$ are estimates of β_0 and $\boldsymbol{\beta}_1$, respectively, and $\widehat{S}(\mathbf{x}_0)$ is an empirical best linear unbiased prediction of $S(\mathbf{x}_0)$. For known covariance structure of S the resulting kriging formula is

$$\widehat{y}(\mathbf{x}_0) = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1^{\mathsf{T}}\mathbf{x}_0 + \widehat{\mathbf{c}}_0^{\mathsf{T}}\mathbf{C}^{-1}(\mathbf{y} - \widehat{\beta}_0 - \widehat{\boldsymbol{\beta}}_1^{\mathsf{T}}\mathbf{x}_0) \tag{6}$$

where

$$\mathbf{C} = [\text{cov}\{S(\mathbf{x}_i), S(\mathbf{x}_j)\}]_{1\leqslant i,j\leqslant n} \quad \text{and} \quad \mathbf{c}_0^{\mathsf{T}} = [\text{cov}\{S(\mathbf{x}_0), S(\mathbf{x}_i)\}]_{1\leqslant i\leqslant n}.$$

The practical implementation of (6) requires a parsimonious model for the interpoint covariances $\text{cov}\{S(\mathbf{x}), S(\mathbf{x}')\}$, $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^2$. Following the recommendations of Stein (1999) we use

$$\text{cov}\{S(\mathbf{x}), S(\mathbf{x}')\} = C_{\boldsymbol{\theta}}(\|\mathbf{x}-\mathbf{x}'\|) \tag{7}$$

where $\|\mathbf{v}\| = \sqrt{\mathbf{v}^{\mathsf{T}}\mathbf{v}}$ and $C_{\boldsymbol{\theta}}$ is member of the Matérn family of covariance functions. It should be pointed out that (7) corresponds to S being *isotropic*, which we view as a reasonable working assumption for the application at hand. The most general such covariance function involves three parameters: $\boldsymbol{\theta} = [\sigma_{\mathbf{x}}^2 \ \rho \ \nu]^{\mathsf{T}}$, where $\sigma_{\mathbf{x}}^2 = \text{Var}\{S(\mathbf{x})\}$ is the variance of the process, ρ is the *range* parameter and controls the distance at which covariances are effectively zero, and ν controls the smoothness of the resulting surface estimate. The full formulation of $C_{\boldsymbol{\theta}}$ is in terms of modified Bessel functions (e.g. Stein, 1999, p. 31) but the special case $\nu = 3/2$ corresponds to

$$C_{\boldsymbol{\theta}}(r) = \sigma_{\mathbf{x}}^2(1+|r|/\rho)e^{-|r|/\rho}. \tag{8}$$

Indeed, in our analysis we work only with this sub-family of the Matérn covariance functions. We chose (8) because it is the simplest member of the Matérn family that results in differentiable surface estimates. We propose to choose ρ via the simple rule

$$\widehat{\rho} = \tfrac{1}{20}\max_{1\leqslant i,j\leqslant n}\|\mathbf{x}_i - \mathbf{x}_j\|. \tag{9}$$

The details behind this choice are given in the Appendix, but the basic idea is to ensure scale invariance and numerical stability. These choices of ν and ρ lead to a reduction from 3 parameters to one to be estimated via (restricted) maximum likelihood (see Section 5). Apart from the obvious reduction in dimensionality, this also allows for use of standard mixed model software for fitting since kriging reduces to a variance component model (see (11) below). Nychka (2000) conjectures that the variance ratio $\sigma_\varepsilon^2/\sigma_{\mathbf{x}}^2$ is much more important than ρ and ν for kriging noisy data. This will be formally investigated in a forthcoming paper by the authors.

Traditionally the $\boldsymbol{\theta}$ in (6) is obtained by variogram analysis of the residuals from the detrending fit $\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1^{\mathsf{T}}\mathbf{x}$, or its quadratic extension, where $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\beta}}_1$ are chosen via least squares (e.g. Venables and Ripley, 1997). As pointed out by O'Connell and Wolfinger (1997), such an approach is quite *ad hoc*. In addition, Stein (1999) raises concerns about variogram estimation. In keeping with the recommendations of O'Connell and Wolfinger (1997) we propose to use a mixed model approach with residual maximum likelihood for estimation of $\boldsymbol{\theta} = \sigma_{\mathbf{x}}^2$. Precedents of this likelihood approach to kriging include Mardia and Marshall (1984) and Zimmerman (1989). However, one is still faced with an $n \times n$ matrix inversion. The Upper Cape

Cod reproductive data involves $n = 1630$ observations, rendering (6) infeasible. An attractive solution is to use *reduced knot* or *low-rank* kriging as proposed by Nychka *et al.*(1997). Let $\{\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_K\}$ be a representative subset of $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ which we will also refer to as knots. This subset can be obtained via an efficient space filling algorithm (e.g. Johnson, Moore and Ylvisaker, 1990; Nychka and Saltzman, 1998). Figure 3 shows the result of applying such an algorithm to the locations in the Upper Cape Cod reproductive data.

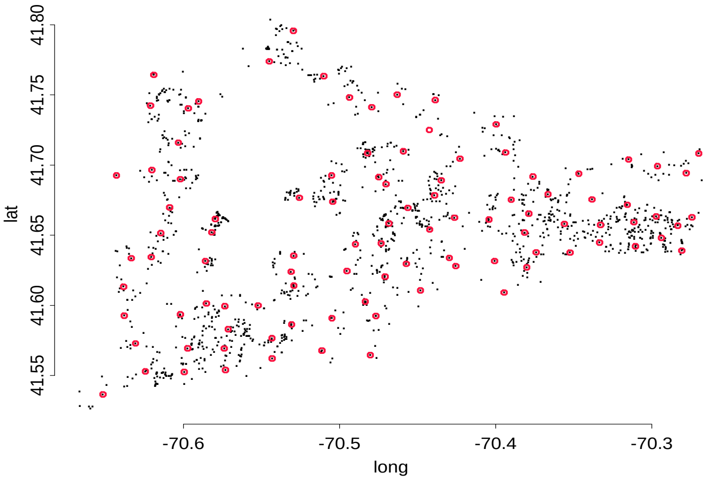


Figure 3: The smaller dots correspond to the geographical locations in the Upper Cape Cod reproductive data. The larger dots correspond to a representative subset of 100 locations for performing low-rank kriging. It was obtained using the space-filling algorithm of Johnson, Moore and Ylvisaker (1990).

Let

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

and

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

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon} \tag{10}$$

where $\text{cov}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \mathbf{I}$, and $\text{cov}(\mathbf{b}) = \sigma_{\mathbf{x}}^2 \boldsymbol{\Omega}^{-1}$. However, for fitting purposes, one should reparameterise to

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \widetilde{\mathbf{Z}}\widetilde{\mathbf{b}} + \boldsymbol{\varepsilon}, \tag{11}$$

where $\widetilde{\mathbf{Z}} = \mathbf{Z}\boldsymbol{\Omega}^{-1/2}$ and $\text{cov}(\widetilde{\mathbf{b}}) = \sigma_{\mathbf{x}}^2 \mathbf{I}$, and utilise the variance component structure. The best linear unbiased prediction corresponding to (10),

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{b}},$$

is nothing more than the set of fitted values on a surface estimate obtained by taking a linear combination of radial basis functions $r_k(\mathbf{x}) = C_0(\|\mathbf{x} - \boldsymbol{\kappa}_k\|/\rho)$, $1 \leqslant k \leqslant$

K, centered about the knots $\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_k$. Indeed, it can be viewed as a member of the class of *Matérn splines* described by Handcock, Meier and Nychka (1994). The popular surface estimation technique known as thin plate splines (e.g. Wahba, 1990; Green and Silverman, 1994) can also be embedded in this framework through the use of generalised covariance functions (e.g. Kitanidis, 1997, p.127). Contributions on kriging/spline equivalences include Kent and Mardia (1994) and Nychka (2000).

In view of (4) and (11) the geoadditive model

$$y_i = \beta_0 + f(s_i) + g(t_i) + \boldsymbol{\beta}_1^{\mathsf{T}}\mathbf{x}_i + S(\mathbf{x}_i) + \varepsilon_i \tag{12}$$

is now trivial to formulate as a single linear mixed model. Put

$$\mathbf{X} = [1\ s_i\ t_i\ \mathbf{x}_i^{\mathsf{T}}]_{1\leqslant i\leqslant n}, \quad \mathbf{Z} = [\mathbf{Z}_s|\mathbf{Z}_t|\mathbf{Z}_{\mathbf{x}}]$$

where $\mathbf{Z}_s$ and $\mathbf{Z}_t$ are defined by (3) and $\mathbf{Z}_{\mathbf{x}} = \widetilde{\mathbf{Z}}$ as given in (11). Then the model has representation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon} \tag{13}$$

where

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}^s \\ \mathbf{b}^t \\ \widetilde{\mathbf{b}} \end{bmatrix} \quad \text{and} \quad \text{cov}\begin{bmatrix} \mathbf{b} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \sigma_s^2\mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_t^2\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_{\mathbf{x}}^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \sigma_\varepsilon^2\mathbf{I} \end{bmatrix}. \tag{14}$$

This is easily implemented using standard mixed model software. Model (12) can be extended to incorporate linear covariates through the $\mathbf{X}\boldsymbol{\beta}$ term. The extension to more than two additive components is straightforward.

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

$$\overline{\mathbf{C}} = [\mathbf{1}|(\mathbf{I} - \tfrac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}})\mathbf{C}_r] \tag{15}$$

rather than $\mathbf{C}$. This convention is adopted in our analysis in Section 7.

5 Amount of smoothing

Penalised spline regression and kriging are both forms of smoothing, and are therefore heavily dependent on the *amount* of smoothing. As mentioned in the previous two sections, the amount of smoothing for both additive components and geostatistical components of a geoadditive model can be quantified through variance component ratios such as $\sigma_\varepsilon^2/\sigma_{\mathbf{x}}^2$. A natural means of choosing the amount of smoothing is to replace variance components with their restricted maximum likelihood (REML) estimates (e.g. Searle, Casella and McCulloch, 1992; O'Connell and Wolfinger, 1997). Since (14) is a simple variance components model, standard mixed model software such as `PROC MIXED` in `SAS` or `lme()` in `S-PLUS` can be called upon to obtain a fully automatic fit.

Even in the additive model context, fully automatic smoothing parameter choice is quite rare. The Markov Chain Monte Carlo approaches of Smith and Kohn (1996), and Shively, Kohn and Wood (1999) produce automatic additive model fits, and the `S-PLUS` function `step.gam()` allows for some automation in smoothing spline-based additive models. However, the more common approach is to use simple rules such as "three degrees of freedom per additive component" (see Section 5.1 below), as is the default for the `gam()` function in `S-PLUS`. Hastie and Tibshirani (1990, pp. 159–161) justify this default by arguing that automatic multiple smoothing parameter selection can be somewhat unstable. This is in keeping with work by, for example, Härdle, Hall and Marron (1988), that raises concerns about the instability of automatic smoothing parameter selection even for single predictor models. Chaudhuri and Marron (1999) recommend looking at curve estimates over a range of smoothing amounts and develop some methodology and graphical devices for doing this systematically.

In summary, while we are attracted by the automatic nature of the mixed model/ REML approach to fitting geoadditive models, we are reluctant to blindly accept whatever answer it provides, and recommend looking at other amounts of smoothing.

5.1 Computation of degrees of freedom

We will now give some details on computation of degrees of freedom values, which are crucial for quantifying the amount of smoothing. For simplicity we restrict description to model (12). Let $\overline{\mathbf{C}}$ be as defined by (15) and let P denote the number of columns in $\overline{\mathbf{C}}$. Then let $\{\mathcal{I}_0,\mathcal{I}_1,\mathcal{I}_2,\mathcal{I}_3\}$ be a partition of the column indices $\{1,\ldots,P\}$ such that $\mathcal{I}_0$ corresponds to the intercept β_0, $\mathcal{I}_1$ and $\mathcal{I}_2$ correspond to $f(s)$ and $g(t)$, and $\mathcal{I}_3$ corresponds to $\boldsymbol{\beta}_1^{\mathsf{T}}\mathbf{x}+S(\mathbf{x})$. For a general matrix $\mathbf{A}$ having P columns define

$$\mathbf{A}_{\mathcal{I}} \equiv \text{sub-matrix of } \mathbf{A} \text{ consisting of columns with indices in } \mathcal{I}.$$

According to this notation

$$\{\overline{\mathbf{C}}_{\mathcal{I}_0},\overline{\mathbf{C}}_{\mathcal{I}_1},\overline{\mathbf{C}}_{\mathcal{I}_2},\overline{\mathbf{C}}_{\mathcal{I}_3}\}$$

represents a partition of the columns of $\overline{\mathbf{C}}$ corresponding to the terms of the additive model (12). Then the degrees of freedom associated with term j, df_j, can be shown to equal

$$\text{df}_j = \text{tr}[\{(\overline{\mathbf{C}}^{\mathsf{T}}\overline{\mathbf{C}})_{\mathcal{I}_j}\}^{\mathsf{T}}\{(\overline{\mathbf{C}}^{\mathsf{T}}\overline{\mathbf{C}}+\sigma_\varepsilon^2\mathbf{B})^{-1}\}_{\mathcal{I}_j}]$$

where

$$\mathbf{B} = \text{diag}\{0, 0, \tfrac{1}{\sigma_s^2}\mathbf{1}_{K_s}, 0, \tfrac{1}{\sigma_t^2}\mathbf{1}_{K_t}, 0, 0, \tfrac{1}{\sigma_{\mathbf{x}}^2}\mathbf{1}_K\}$$

and $\mathbf{1}_p$ denotes a p-dimensional vector of ones.

6 Inference

6.1 Variability bands

Variability bands in function estimation are usually obtained by adding and subtracting twice the estimated standard error of the estimated function (e.g. Bowman and Azzalini, 1997, pp. 75–76). Bias aside, they can be interpreted as approximate pointwise confidence intervals (Hastie and Tibshirani, 1990). They are also useful for detection of leverage and display of inherent variability. For additive models and geoadditive models in the linear mixed model framework the standard errors are easily derived using standard multivariate statistical manipulations after obtaining an estimate of $\text{Cov}([\widehat{\boldsymbol{\beta}}^{\mathsf{T}} \widehat{\mathbf{b}}^{\mathsf{T}}]^{\mathsf{T}} | \mathbf{b})$.

6.2 Hypothesis tests

Another advantage of the mixed model framework is that tests of hypotheses can be performed within the likelihood ratio paradigm. For a general statistical model with data vector $\mathbf{y}$ and parameter vector $\boldsymbol{\theta}$ the test statistic is

$$-2\log\{\text{LR}(\mathbf{y})\} = -2\{\ell(\widehat{\boldsymbol{\theta}}_0; \mathbf{y}) - \ell(\widehat{\boldsymbol{\theta}}; \mathbf{y})\} \tag{16}$$

where $\widehat{\boldsymbol{\theta}}_0$ and $\widehat{\boldsymbol{\theta}}$ are the maximum likelihood estimates of $\boldsymbol{\theta}$ under H_0 and H_1, respectively, and $\ell(\boldsymbol{\theta}; \mathbf{y})$ is the log-likelihood. Under the assumption of normal errors (16) is easy to compute using standard mixed model software. For example, in (1), linearity of the effect of s can be assessed through a test of the hypotheses

$$\begin{aligned} H_0 &: \sigma_s^2 = 0 \\ H_1 &: \sigma_s^2 > 0. \end{aligned}$$

The overall effect of s can be assessed through a test of the hypotheses

$$\begin{aligned} H_0 &: \beta_1 = \sigma_s^2 = 0 \\ H_1 &: \beta_1 \neq 0 \text{ or } \sigma_s^2 > 0. \end{aligned} \tag{17}$$

A forthcoming paper by Aerts, Claeskens, Ruppert and Wand (2000) will provide an in-depth investigation into such tests in the additive model context.

7 Analysis of Upper Cape Cod reproductive data

The geoadditive model described in Section 4 was implemented using the S-PLUS function lme(), corresponding to Version 2.1 of the NLME module. The largest geoaddative model required for analysis of the Upper Cape Cod data took $1\frac{1}{2}$ minutes to run on our workstations. This is quite fast considering the sophistication of the model and the fact that the smoothing parameter choice is automatic.

We first analysed the data using fully automatic smoothing parameter choice based on REML. Model selection for the non-linear components was performed using likelihood ratio statistics as described in Section 6.2, while the linear components were chosen according to the approximate Z-value given by lme(). Residuals from final model fits were checked, and showed no discernible patterns. Table 2 summarises the results for the selected model. A summary of the likelihood ratio statistics for non-linear effects is given in Table 3. Although the null distribution for $-2\log\{\text{LR}(\mathbf{y})\}$ has some complications that does not yet allow us to report p-values

Table 2: Summary of final REML-based fit of geoadditive model for Upper Cape Cod reproductive data.

	birthweight		gestational age	
	coef	p-value	coef	p-value
male	162.78	0.0000		
maternal age	-7.34	0.0067		
preg. hyperten.	-189.40	0.0472		
light prev. birth	-442.20	0.0009		
heavy prev. birth	306.52	0.0018		
renal disease	-640.25	0.0552		
black	-148.30	0.0271		
asian	-219.91	0.0515		
drinks per week	-42.34	0.0102		
plurality	-845.30	0.0000	-2.6308	0.0000
uterine bleeding	-412.51	0.0097	-1.3856	0.0418
psychiatric	-525.53	0.0259	-2.0454	0.0430
incomp. cervix	-931.26	0.0485	-3.0750	0.0313
eclampsia	-1073.60	0.0226	-5.4784	0.0066
cig's per day			-0.0249	0.0125
	df		df	
parity	2.780			
cig's per day	2.326			
years of education	3.671			
prenatal visits	2.259		3.331	
maternal age			3.526	
longitude,latitude	2.018		2.006	

(Aerts *et al.*2000) the magnitude of the statistics suggests, in most cases, a very high degree of statistical significance.

Figure 4 displays all non-linear covariate effects. While our primary concern in this study is geographical effects on reproductive outcomes, the non-linear covariate effects depicted here are quite interesting in their own right.

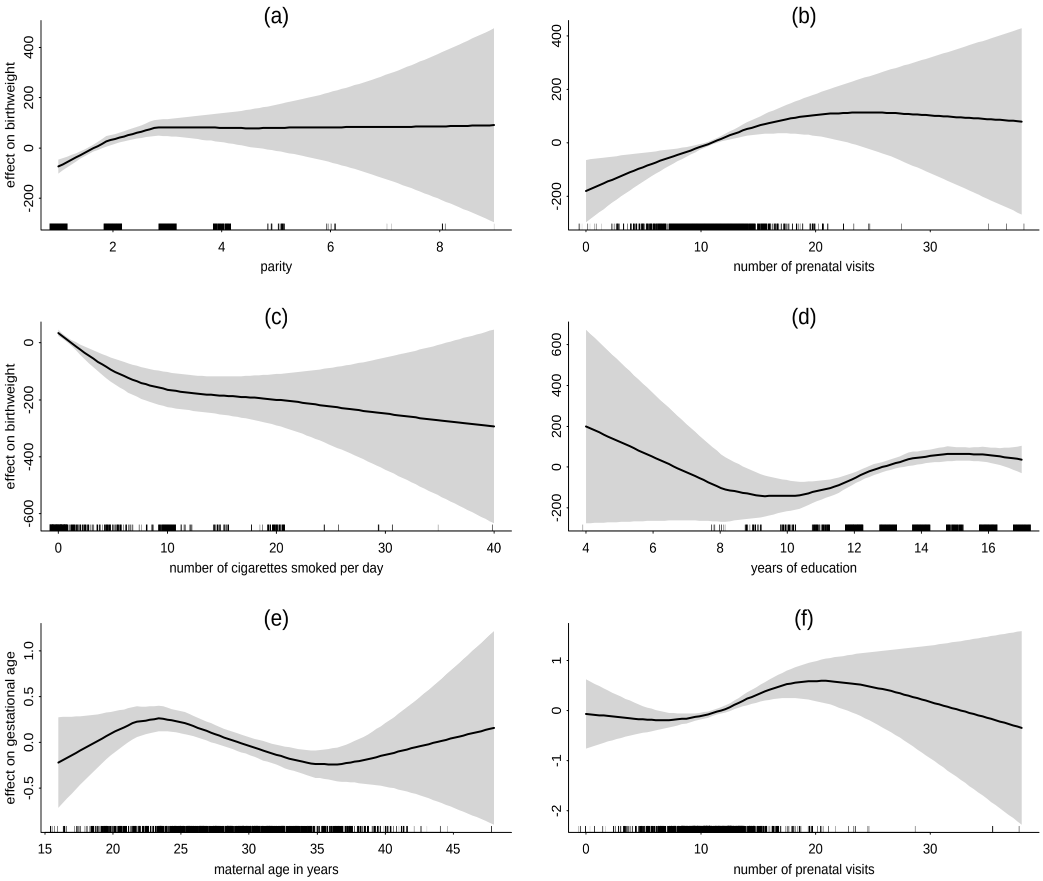


Figure 4: Nonlinear terms from *geoadditive* model fit. Panels (a)-(d) correspond to birthweight. Panels (e)-(f) correspond to gestational age.

Most importantly, the geographical component is *not* found to be significant based on REML variance component estimation. As seen in Table 2 REML chooses only 2.018 degrees of freedom for location for prediction of birthweight, and 2.006 degrees of freedom for prediction of gestational age. This effectively corresponds to a planar fit. The likelihood ratio statistics for non-linearity of the geographical components were both very small. When the model was re-fit with longitude and latitude as linear effects the p-values were large for both birthweight and gestational age.

In spite of the REML-based analysis showing no geographical effect, we obtained fits where a range of higher degrees of freedom values were used for the kriging component. The degrees of freedom values for the other non-linear components were fixed at their REML values. The results are shown in Figures 5 and 6. Meaningful geographical variation is not discernible for gestational age. However, Figure 5 suggests some regions with lower than average mean birthweight; particularly in the north-western strip. The Massachusetts Military Reservation is directly east of this strip, and it has long been identified as a source of contamination. While this result is exploratory, rather than confirmatory, it does suggest the possibility of a link between low birthweights and proximity to the military reservation; and may warrant further investigation.

Table 3: Likelihood ratio statistics for non-linear terms.

	birthweight	gestational age
null hypothesis	$-2\log\{\text{LR}(\mathbf{y})\}$	$-2\log\{\text{LR}(\mathbf{y})\}$
effect of parity	55.912	
effect of cig's per day	63.333	
effect of years of education	46.298	
effect of prenatal visits		7.265
effect of maternal age		2.228
linearity of parity	30.686	
linearity of cig's per day	27.395	
linearity of years of education	27.487	
linearity of prenatal visits	25.194	4.840
linearity of maternal age		4.079

Figures 5 and 6 are in keeping with the recommendations of Chaudhuri and Marron (1999) who provide some convincing arguments for looking at smooths across several values of the smoothing parameter, not just that one chosen via an automatic method. These authors develop a graphical device, named *SiZer*, to facilitate the problem of testing for features in a function, while recognising the inherent dependence on the amount of smoothing in the function estimate. Bivariate extensions have been recently developed (Godtliebsen, Marron and Chaudhuri, 2000a, 2000b). An interesting future project would be a SiZer-type analysis of these data to systematically assess the presence of any 'hot spots', after accounting for covariate effects.

An S-PLUS module tailored to fitting geoadditive models has been developed by the authors and is available on request. (The current e-mail address of the second author is mwand@hsph.harvard.edu.)

8 Generalised geoadditive models

The reproductive outcomes birthweight and gestational age are continuous and free of any significant skewness, so the Gaussian mixed model is an adequate vehicle for the analysis of that data. In the case where the response is categorical (e.g. a binary or count variable) or heavily skewed, *generalised* linear mixed models need to be used instead. We might call the result *generalised geoadditive models*.

Given the earlier sections, generalised geoadditive models are straightforward to formulate. For example, if the response y is binary then the analogue of (12) is

$$\text{logit}\{P(y_i = 1|S)\} = \beta_0 + f(s_i) + g(t_i) + \boldsymbol{\beta}_1^\mathsf{T}\mathbf{x}_i + S(\mathbf{x}_i)$$

and this can be fit through a mixed model of the form

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

where $\mathbf{b}$ is a random effects vector with covariance structure exemplified by that given in (14). In the case where all covariate effects are linear (18) essentially corresponds to the model proposed by Diggle, Tawn and Moyeed (1998).

The fitting of such models using maximum likelihood is quite complicated due to the presence of intractable integrals in the likelihood. Nevertheless, there has been a great deal of research on the topic since the early 1990s (e.g. Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Zeger and Karim, 1993; Lin and Breslow, 1997; McCulloch, 1997; Diggle, Tawn and Moyeed, 1998; Booth and Hobert, 1999) and, in theory, any of these approaches can be used to fit generalised geoadditive models. Future research will investigate the practicalities in the context of geoadditive models.

9 Closing remarks

The geoadditive model is an effective vehicle for the analysis of spatial epidemiologic data and other applications where geographic point data are accompanied by covariate measurements. The low-rank mixed model formulation allows for straightforward implementation and fast processing of large data bases, thus fascilitating use of the model in surveillance of disease clusters.

The geoadditive model has been shown to be useful for analysis of the Upper Cape Cod reproductive data. It properly accounts for all covariate information before producing disease maps. In the case of gestational age it has been seen that no residual geographical effect is present. The birthweight analysis is slightly suggestive, but geographical variation cannot yet be concluded.

Acknowledgements

We are grateful to Bob Knorr, Joel Schwartz, Mike Wright and the Department of Public Health, Commonwealth of Massachusetts, for making the data available for this analysis. The paper has benefited from discussions with Jonathan French, Michael O'Connell, Doug Nychka, Louise Ryan and Lance Waller. The research of E. E. Kammann was supported by NIH grant T32 ES07142-18.

Appendix A

Derivation of (9) for choice of range parameter

Let $D = \max_{1 \leqslant i,j \leqslant n} \|\mathbf{x}_i - \mathbf{x}_j\|$ denote the maximal inter-point distance in the set $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^2$. If an interval is drawn between the two most distant points then we ask that a radial basis function centred on the midpoint of that interval have approximate support over half of the interval. The 'half' here is simply a nominal value chosen to ensure numerical stability. For 'approximate support' we nominally choose the edge of support to correspond to the function equaling approximately 5% of the height at the peak. For covariance function (9) this leads to the equation

$$e^{-D/(4\rho)}\{1 + D/(4\rho)\} = 0.05.$$

The solution is $\rho = D/18.9755...$ which we round off to $\rho = D/20$.

References

Aerts, M., Claeskens, G., Ruppert, D. and Wand, M.P. (2000). Likelihood ratio testing in penalized spline additive models. Unpublished manuscript.

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 Statistics Society, Series B* , **61**, 265–285.

Bowman, A.W. (1998). Comment on paper by Diggle, Tawn and Moyeed. *Applied Statistics*, **47**, 334.

Bowman, A.W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*, Oxford: Clarendon Press.

Breslow, N.E. and Clayton, D.G. (1993). Approximated inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

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.

Chaudhuri, P. and Marron, J.S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, **94**, 807–823.

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

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998). Model-based geostatistics (with discussion). *Applied Statistics*, **47**, 299–350.

Durbán Reguera M.L. (1998). Modelling spatial trends and local competition effects using semiparametric additive models. *PhD Thesis, Heriot-Watt University*.

Durbán, M., Hackett, C., Currie, I. and Newton, A. (2000). Analysis of spatial trends in field trials using semiparametric additive models. Unpublished manuscript.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, **89**, 89–121.

Gawande, A. (1999). The cancer-cluster myth. *The New Yorker*, Feb. 8, 34–37.

Godtliebsen, F., Marron, J.S. and Chaudhuri, P. (2000a). Significance in scale space. Unpublished manuscript.

Godtliebsen, F., Marron, J.S. and Chaudhuri, P. (2000b). Significance in scale space for density estimation. Unpublished manuscript.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

Handcock, M.S., Meier, K. and Nychka, D. (1994). Comment on paper by Laslett. *Journal of the American Statistical Association*, **89**, 401–403.

Härdle, W., Hall, P. and Marron, J.S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association* , **83**, 86–101.

Hastie, T.J. (1996). Pseudosplines. *Journal of the Royal Statistical Society, Series B*, **58**, 379–396.

Hastie, T.J. and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Hobert, J. P., Altman, N. S. and Schofield, C. L. (1997). Analyses of fish species richness with spatial covariate *Journal of the American Statistical Association*, **92**, 846–854.

Johnson, M.E., Moore, L.M. and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, **26**, 131–148.

Kelsall, J.E. and Diggle, P.J. (1998). Spatial variation in risk of disease: A nonparametric binary regression approach. *Applied Statistics*, **47**, 559-573.

Kent, J.T. and Mardia, K.V. (1994). The link between kriging and thin plate splines. In *Probability, Statistics and Optimization: a Tribute to Peter Whittle* (ed. F.P. Kelly), pp. 325–339. Chichester: Wiley.

Kitanidis, P.K. (1997). *Introduction to Geostatistics*, Cambridge: Cambridge University Press.

Lin, X. and Breslow, N.E. (1997). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, **91**, 1007–1016.

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

Mardia, K.V. and Marshall, R.J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **72**, 135–146.

Marx, B.D. & Eilers, P.H.C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis*, **28**, 193–209.

Nychka, D.W. (2000). Spatial process estimates as smoothers. In *Smoothing and Regression* (M. Schimek, ed.), Heidelberg: Springer-Verlag.

Nychka, D., Bailey, B., Ellner, S., Haaland, P. and O'Connell, M. (1997). *FUNFITS: Data Analysis and Statistical Tools for Estimating Functions.* Unpublished manuscript available at `www.stat.ncsu.edu/`$\sim$`nychka/man.html`

Nychka, D. and Cummins, D.J.(1996). Comment on paper by Eilers and Marx. *Statistical Science*, **11**, 104–105.

Nychka, D. and Saltzman, N. (1998). Design of Air Quality Monitoring Networks. in Case Studies in Environmental Statistics Nychka, D., Cox, L., Piegorsch, W. ed., Lecture Notes in Statistics, Springer-Verlag.

O'Connell, M.A. and Wolfinger, R.D. (1997). Spatial regression models, response surfaces, and process optimization. *Journal of Computational and Graphical Statistics*, **6**, 224–241.

Ruppert, D. (2000). Selecting the number of knots for penalized splines. Unpublished manuscript.

Ruppert, D. and Carroll, R.J. (2000). Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics*, **42**, 205–224.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance Components*, New York: Wiley.

Shively, T.S., Kohn, R. and Wood, S. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior. *Journal of the American Statistical Association*, **94**, 777–794.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics*, **75**, 317–344.

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

Venables, W.N. and Ripley, B.D. (1997). *Modern Applied Statistics with S-PLUS*. New York: Springer.

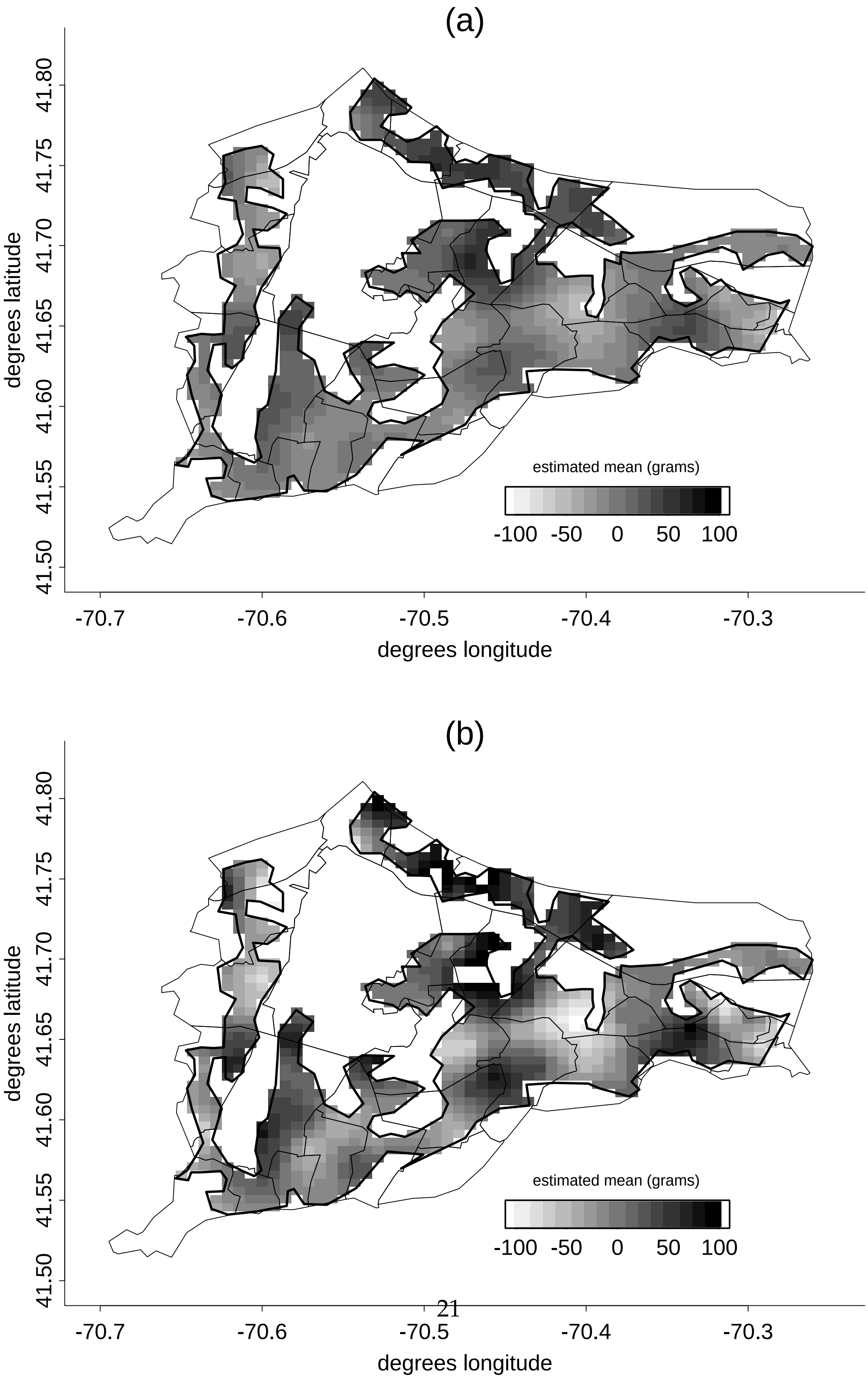
Wahba, G. (1990). *Spline Models for Observational Data., Philadelphia: SIAM.*

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, **48**, 233–243.

Zeger, S.L. and Karim, M. R. (1993). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.

Zimmerman, D. (1989). Computationally efficient restricted maximum likelihood estimation of generalized covariance functions. *Mathematical Geology*, **21**, 655–672.

Figure 5: Geographical components of geoadditive model fits to birthweight with user specified degrees of freedom value: (a) 20 degrees of freedom, (b) 40 degrees of freedom. The light lines correspond to census block groups.



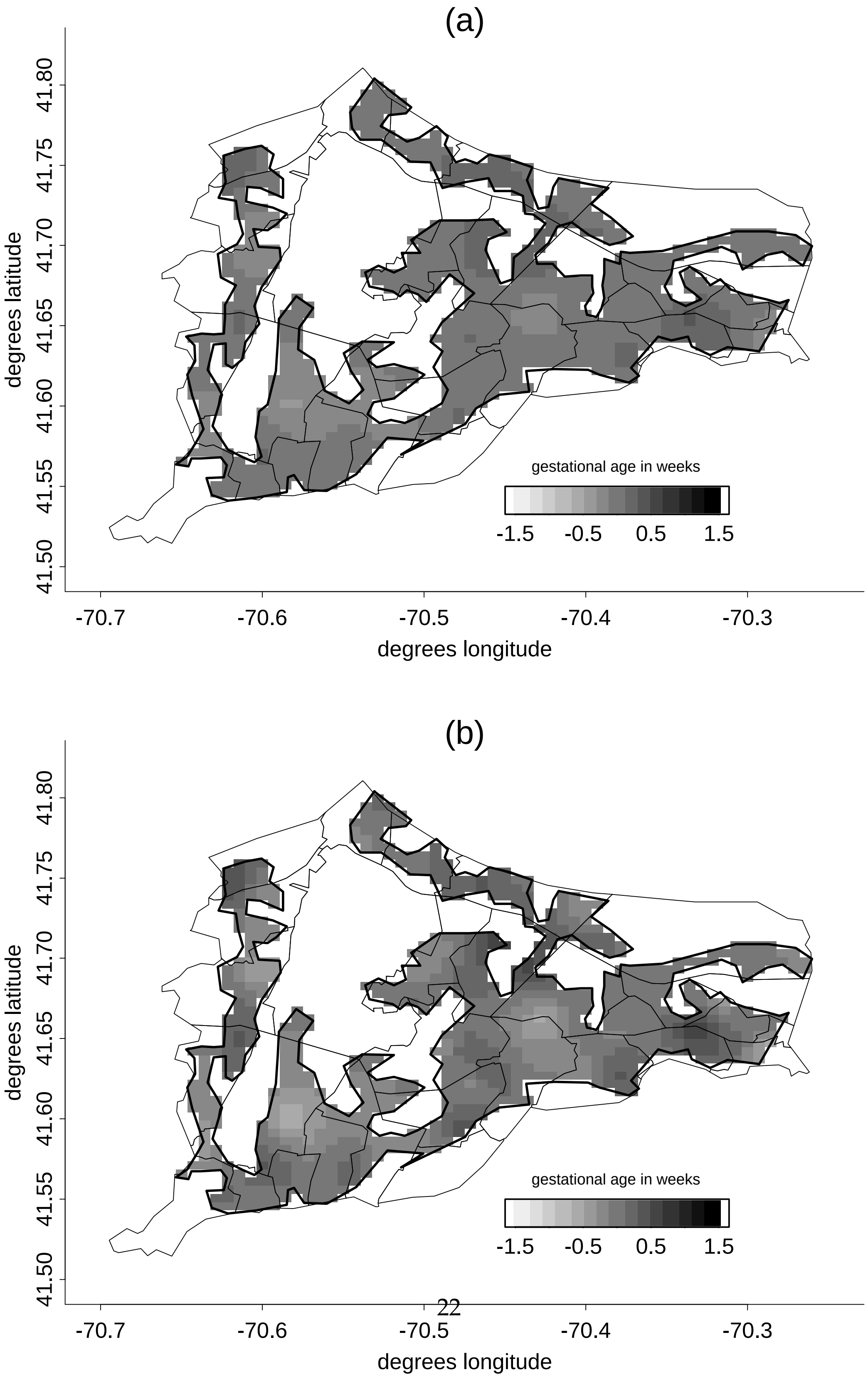


Figure 6: Geographical components of geoadditive model fits to gestational age with user specified degrees of freedom value: (a) 20 degrees of freedom, (b) 40 degrees of freedom.The light lines correspond to census block groups.