Appl. Statist. (1998)
47, *Part* 4, *pp.* 559–573

Spatial variation in risk of disease: a nonparametric binary regression approach

Julia E. Kelsall† and Peter J. Diggle

Lancaster University, UK

[Received December 1996. Revised January 1998]

Summary. A common problem in environmental epidemiology is the estimation and mapping of spatial variation in disease risk. In this paper we analyse data from the Walsall District Health Authority, UK, concerning the spatial distributions of cancer cases compared with controls sampled from the population register. We formulate the risk estimation problem as a nonparametric binary regression problem and consider two different methods of estimation. The first uses a standard kernel method with a cross-validation criterion for choosing the associated bandwidth parameter. The second uses the framework of the generalized additive model (GAM) which has the advantage that it can allow for additional explanatory variables, but is computationally more demanding. For the Walsall data, we obtain similar results using either the kernel method with controls stratified by age and sex to match the age–sex distribution of the cases or the GAM method with random controls but incorporating age and sex as additional explanatory variables. For cancers of the lung or stomach, the analysis shows highly statistically significant spatial variation in risk. For the less common cancers of the pancreas, the spatial variation in risk is not statistically significant.

Keywords: Binary regression; Cross-validation; Epidemiology; Generalized additive models; Kernel smoothing

1. Introduction

Suppose that in a geographical region A we are given the locations of all cases of a particular disease. A natural question to ask is whether or not the disease risk varies spatially. If there is some evidence of spatial variation, then knowledge of the characteristics of particular subregions in which risk appears to be higher than average may lead to new hypotheses regarding possible causal mechanisms for the disease. However, the locations of cases are not enough to answer this question, since they will at least in part reflect the spatial distribution of the population at risk in the region. A sensible approach is then to compare the spatial distribution of the cases with that of a set of carefully selected controls from the population at risk. Bithell (1990, 1992) and Kelsall and Diggle (1995a, b) approached this problem as one of density ratio estimation. Lawson and Williams (1993) also described a method which they called 'extraction mapping' based on kernel regression. In this paper we consider the use of nonparametric binary regression, which we find to be more flexible than the density ratio approach; using generalized additive model (GAM) methodology (Hastie and Tibshirani, 1990) we can estimate the effects of other covariates in addition to spatial location.

Consider the sets of points consisting of the locations of cases and controls as observations from two Poisson processes I and II on the region $A \subset \mathbb{R}^2$, with intensities $\lambda_1(x)$ and $\lambda_2(x)$ respectively. The log-risk function, apart from an additive constant, is

†*Address for correspondence*: Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, UK.
E-mail: julia.kelsall@lancaster.ac.uk

 0035–9254/98/47559

$$\rho(x) = \log\{\lambda_1(x)/\lambda_2(x)\},$$

and we wish to investigate the spatial variation of $\rho(x)$ on A.

This work was motivated by data consisting of the residential locations, age and sex of individuals who died of one of various types of cancer in the region of the Walsall District Health Authority between 1982 and 1992. Methods will be developed to investigate whether there is any spatial variation in risk of cancers of the lung, stomach and pancreas over the region.

In the next section we briefly review the density ratio estimation approach and introduce our two methods — kernel binary regression and the GAM method. For each, we describe how to choose the associated bandwidth or smoothing parameter, whose numerical value determines the degree of smoothing to be applied to the data. Section 3 reports the results of a small simulation study to compare the methods. In Section 4 we describe a Monte Carlo method for assessing the uncertainty associated with the estimates; we obtain tolerance contours which enable the identification of areas of unusually high or low risk, and we provide a global test of the null hypothesis of constant risk over the region. The methodology is applied to the cancer mortality data of the Walsall District Health Authority in Section 5. Section 6 briefly discusses the relative merits of the alternative methods.

2. Estimation methods

In practice, it is not possible to ensure complete ascertainment of cases in studies of this kind. For example, in the cancer mortality data, some cases may have been missed owing to inaccurate recording of the cause of death. For our risk estimates to be valid, we need to assume that non-ascertainment of cases is spatially random, i.e. that the cases available for analysis are an independent random sample from the totality of cases. We therefore suppose that we observe unknown proportions q_1 and q_2 of those points which constitute partial realizations of the two independent Poisson point processes I and II on the region $A \subset \mathbb{R}^2$. Denote by x_i, $i = 1, \ldots, n_1$, the n_1 observed points from process I and by x_i, $i = n_1 + 1, \ldots, n_1 + n_2$, the n_2 observed points from process II. We first consider the density ratio approach and then introduce the two binary regression alternatives.

2.1. Density ratio

We give a brief review of the density ratio approach of Kelsall and Diggle (1995a, b). Conditionally on the values of n_1 and n_2, the data can be regarded as a pair of independent random samples from probability distributions with densities $f_1(x)$ and $f_2(x)$ such that

$$f_j(x) = \alpha_j^{-1}\lambda_j(x) \qquad \text{with } \alpha_j = \int_A \lambda_j(x)\,\mathrm{d}x,$$

for $j = 1, 2$. Letting $r(x)$ be the log-ratio of densities, we find that

$$r(x) = \log\{f_1(x)/f_2(x)\} = \rho(x) - c_1,$$

where $c_1 = \log(\alpha_1/\alpha_2)$.

An estimator for $r(x)$ is thus

$$\hat{r}_h(x) = \log\{\hat{f}_{1h}(x)/\hat{f}_{2h}(x)\},$$

where $\hat{f}_{1h}(x)$ and $\hat{f}_{2h}(x)$ are kernel estimators of $f_1(x)$ and $f_2(x)$ respectively. We use a common

bandwidth h for the numerator and denominator which eliminates the bias of $\hat{r}_h(x)$ when $f_1(x) = f_2(x)$. Thus,

$$\hat{f}_{jh}(x) = n_j^{-1} \sum_{i=1}^{n_j} K_h(x - x_i),$$

where $K_h(u) = h^{-2} K(h^{-1}u)$ and K is a radially symmetric kernel function. In what follows, we use the standard bivariate normal kernel, $K(u) = (2\pi)^{-1} \exp(-\frac{1}{2}\|u\|^2)$.

A cross-validation method for choosing the bandwidth with the aim of minimizing $\int \{\hat{r}_h(x) - r(x)\}^2 \, dx$ was devised using Taylor series expansion arguments (Kelsall and Diggle, 1995a) and is based on the score

$$\mathrm{CV}_1(h) = -\int_I \hat{r}_h(x)^2 \, dx - 2n_1^{-1} \sum_{i=1}^{n_1} \hat{r}_h^{-i}(x_i)/\hat{f}_{1h}^{-i}(x_i) + 2n_2^{-1} \sum_{i=n_1+1}^{n_1+n_2} \hat{r}_h^{-i}(x_i)/\hat{f}_{2h}^{-i}(x_i).$$

The superscript of $-i$ denotes estimation using all the data except x_i, usually called *leave-one-out* estimation. We choose the value of h which minimizes $\mathrm{CV}_1(h)$. Edge-corrected versions of the estimates must be used, and performing these edge corrections is a non-trivial operation.

2.2. Binary regression

Simpler cross-validation methods can be obtained by reformulating the problem as one of binary regression. We attach binary labels $y_1, \ldots, y_n$ to the points $x_1, \ldots, x_n$ such that

$$y_i = \begin{cases} 1 & \text{if } x_i \text{ is from group I,} \\ 0 & \text{otherwise.} \end{cases}$$

Conditional on the points x_i, the y_i are realizations of mutually independent Bernoulli random variables Y_i with $P(Y_i = 1 | X_i = x) = p(x)$, where

$$p(x) = \frac{q_1 \, \lambda_1(x)}{q_1 \, \lambda_1(x) + q_2 \, \lambda_2(x)}.$$

It follows that

$$\mathrm{logit}\{p(x)\} = \rho(x) + c_2,$$

where $c_2 = \log(q_1/q_2)$.

Apart from an additive constant, both $r(x)$ and $\mathrm{logit}\{p(x)\}$ are equal to $\rho(x)$. This means that, from the point of view of estimating spatial variation in risk, we can approach the problem either via nonparametric regression of the y_i against the x_i or via density ratio estimation.

Overviews of nonparametric regression techniques can be found in Härdle (1990), Green and Silverman (1994) and Wand and Jones (1995). These include kernel methods, smoothing splines and local likelihood methods, among which local logistic regression (Fan *et al.*, 1995) is specifically intended for binary response variables. We here consider the simpler Nadaraya–Watson kernel regression estimator, first suggested in the binary setting by Copas (1983). This estimates $p(x)$ as

$$\hat{p}_h(x) = \sum_{i=1}^{n} K_h(x - x_i) \, y_i \Big/ \sum_{i=1}^{n} K_h(x - x_i),$$

where the kernel function $K_h(\cdot)$ and bandwidth h are defined in Section 2.1. Using this estimator, straightforward algebraic manipulation shows that

$$\text{logit}\{\hat{p}_h(x)\} = \hat{r}_h(x) + c_3,$$

where $c_3 = \log(n_1/n_2)$. The kernel regression approach therefore leads to the same estimate of $\rho(x)$ as the kernel density ratio approach, except for an additive constant. However, the new approach suggests different cross-validation methods for choosing h in practice.

2.3. Bandwidth selection

2.3.1. Likelihood cross-validation

This method of choosing a bandwidth is motivated by the maximum likelihood principle. Considering the data in the binary regression framework, we can write the likelihood for the probability function $p(\cdot)$ as

$$L\{p(\cdot)\} = \prod_{i=1}^{n} p(x_i)^{y_i}\{1 - p(x_i)\}^{1-y_i}.$$

If $p(\cdot)$ were specified by a parametric model, then the maximum likelihood principle would lead to estimating the parameters by maximizing $L\{p(\cdot)\}$. In the nonparametric setting, this would lead to the unsatisfactory estimates $p(x_i) = 1$ or $p(x_i) = 0$ according to whether $Y_i = 1$ or $Y_i = 0$ respectively. However, we can use a cross-validatory version of the maximum likelihood principle if we estimate $p(\cdot)$ to minimize

$$\text{CV}_2(h) = \left[\prod_{i=1}^{n} \hat{p}_h^{-i}(x_i)^{y_i}\{1 - \hat{p}_h^{-i}(x_i)\}^{1-y_i}\right]^{-1/n} \tag{2.1}$$

with respect to h, where $\hat{p}_h^{-i}(x_i)$ is the leave-one-out estimate of $p(x)$ evaluated at x_i. In the definition of $\text{CV}_2(h)$, we raise the 'likelihood' to the power of $-1/n$ to make the value of the criterion less sensitive to the sample size n and to convert it to a minimization problem. This form of cross-validation has also been suggested by Azzalini *et al.* (1989).

2.3.2. Least squares cross-validation

One of the standard ways of choosing a smoothing parameter in nonparametric regression is to use least squares cross-validation. This is known to work effectively for non-binary data (Härdle, 1990). The cross-validation function to be minimized is

$$\text{CV}_3(h) = n^{-1}\sum_{i=1}^{n}\{y_i - \hat{p}_h^{-i}(x_i)\}^2, \tag{2.2}$$

where the y_i are binary labels. As with the likelihood-based method we use a leave-one-out procedure to avoid an interpolating solution.

2.3.3. Weighted least squares cross-validation

As we are dealing with binary responses Y_i, we know that $\text{var}(Y_i|X_i = x_i) = p(x_i)\{1 - p(x_i)\}$. This makes an ordinary least squares criterion less natural than it would be for responses with a constant variance and suggests instead a new cross-validation criterion. For the new criterion, we obtain a preliminary choice of smoothing parameter, h_0 say, by the ordinary least squares cross-validation of the previous section, and we use this to estimate the variance

function. We can then choose a bandwidth h which minimizes a weighted version of function (2.2), with the inverse of the estimated variance as weights. This gives the criterion of minimizing

$$\mathrm{CV}_4(h) = n^{-1} \sum_{i=1}^{n} \frac{\{y_i - \hat{p}_h^{-i}(x_i)\}^2}{\hat{p}_{h_0}^{-i}(x_i)\{1 - \hat{p}_{h_0}^{-i}(x_i)\}}. \tag{2.3}$$

If we replaced h_0 by h, the denominator would no longer simply give weights to the data values as desired but would take an active role in the minimization of expression (2.3) and bias the minimization towards unreasonably large values of h. This is connected with the well-known analogous phenomenon for parametric regression models: for response variables Y_i with expectations $\mu_i(\theta)$ and variances $v_i(\theta)$, minimization of $\Sigma\{y_i - \mu_i(\theta)\}^2/v_i(\theta)$ is equivalent to using a biased estimating equation for θ (see, for example, McCullagh (1983) and Barry *et al.* (1997)).

2.4. Generalized additive models

A natural extension of the binary regression approach is to consider including covariate terms in the regression. This leads us to the GAM approach (Hastie and Tibshirani, 1990).

A GAM can be thought of as a generalized linear model (GLM) (see McCullagh and Nelder (1989)), which has been extended to include arbitrary smooth functions in addition to linear terms in the linear predictor. In our context, we consider a binary response variable Y with associated $P(Y = 1) = p(x, u)$ depending on the explanatory variables u and spatial location x. A GAM with a logit link function would then take the form

$$\mathrm{logit}\{p(x, u)\} = u'\beta + g(x) \tag{2.4}$$

where $\beta = (\beta_1, \ldots, \beta_r)$, say, and the only assumption about g is that it is a *smooth* function of x. GAMs are fitted by an iteratively weighted additive model procedure that is an extended version of the iteratively weighted least squares of GLMs. Following Hastie and Tibshirani (1990), estimation for the semiparametric model (2.4) using kernel regression for data (y_i, x_i, u_i), $i = 1, \ldots, n$, proceeds according to the following algorithm.

(a) Set $\hat{g}(x) = 0$; set $\hat{\beta}$ to be the maximum likelihood estimate on fitting the GLM with

$$\mathrm{logit}\{p(x, u)\} = u'\beta.$$

(b) Set $\hat{\eta}_i = u_i'\hat{\beta} + \hat{g}(x_i)$ and $\hat{p}_i = \exp(\hat{\eta}_i)/\{1 + \exp(\hat{\eta}_i)\}$.

(c) For $i = 1, \ldots, n$, construct the *adjusted dependent variable*

$$z_i = \hat{\eta}_i + \frac{y_i - \hat{p}_i}{\hat{p}_i(1 - \hat{p}_i)}$$

and weights $w_i = \hat{p}_i(1 - \hat{p}_i)$.

(d) Fit a weighted additive model of the form $Z = u'\beta + g(x) + \epsilon$, using kernel regression with weights w_i as follows.

(i) Put $s_i = z_i - u_i'\hat{\beta}$ and perform a weighted kernel regression of s_i on x_i with weights w_i:

$$\hat{g}(x) = \sum_{i=1}^{n} w_i K_h(x - x_i) S_i \Big/ \sum_{i=1}^{n} w_i K_h(x - x_i).$$

(ii) Regress the $z_i - \hat{g}(x_i)$ on the u_i-values by weighted least squares with weights w_i to obtain new estimates $\hat{\beta}$.
(iii) Repeat steps (i) and (ii) until the estimates converge.

(e) Repeat steps (b)–(d) until the estimates converge.

A special case is when there are no covariates u, so that

$$\text{logit}\{p(x)\} = g(x).$$

Although the major attraction of the GAM approach lies in its ability to incorporate covariate adjustments, the special case is of interest because its performance can be directly compared with the density ratio and kernel regression approaches. Note that there is no direct algebraic equivalence between the GAM estimate of $p(x)$ and the corresponding kernel estimate computed from the same data, although the underlying models are equivalent.

This method also requires a bandwidth to be chosen. The likelihood and least squares methods of kernel regression are in principle applicable here, but in practice they are computationally infeasible because of the large number of iterations involved. We shall therefore use a form of weighted least squares cross-validation for the nonparametric regression step of each iteration. At each smoothing step (i), we choose the value of h which minimizes

$$\text{CV}_5(h) = n^{-1} \sum_{i=1}^{n} w_i \{z_i - \hat{g}^{-i}(x_i)\}^2, \tag{2.5}$$

where $\hat{g}^{-i}(x_i)$ is the estimate of $g(x_i)$ constructed with bandwidth h using all except the data pair (x_i, z_i).

3. Simulations

In this section we use a small simulation study to compare one-dimensional versions of the methods outlined in the previous section. For a more extensive simulation study see Kelsall (1996). We consider five methods for estimating $\rho(x)$ up to a constant:

(a) density ratio estimation with the associated cross-validation method;
(b) binary kernel regression with the likelihood-based cross-validation method;
(c) binary kernel regression with the least squares cross-validation method;
(d) binary kernel regression with the weighted least squares cross-validation method;
(e) GAM regression with the associated weighted least squares cross-validation method.

For each simulation, we consider the interval $I = (0, 1)$ and fix the relative risk, $\exp\{r(x)\} = f_1(x)/f_2(x)$, the control density $f_2(x)$ and the sample sizes n_1 and n_2. Without loss of generality, assume that $\rho(x) = r(x)$. We generate two sets of points from densities $f_2(x)$ and

$$f_1(x) = \exp\{r(x)\} f_2(x)$$

of sizes n_1 and n_2, and obtain an estimate $\hat{\rho}(x)$ using each of the five methods. A measure of the error of this estimate is taken as a version of the *integrated squared error* (ISE),

$$\text{ISE} = \min_c \left[\int_I \{\hat{\rho}(x) + c - \rho(x)\}^2 \, \text{d}x \right],$$

in which we minimize over c because $\rho(x)$ is only estimated up to an additive constant. We use

$$\exp\{r(x)\} = 1 + 0.75 \, \sin(6\pi x)$$

as the risk function, representing a highly fluctuating risk, and combine this with three control densities:

(a) $f_2(x) = 1$;
(b) $f_2(x) = 1 + 0.7 \cos\{2\pi(x - 0.5)\}$;
(c) $f_2(x) = 1 + 0.8 \cos\{6\pi(x - 0.5)\}$.

The first of these represents a constant control intensity, which is highly unlikely to be found in practice. The other two represent more likely fluctuating densities.

To automate this procedure, for each combination we calculate the cross-validation function for 15 values of h, regularly spaced on a log-scale between 0.005 and 2, and choose the value for h which gives the smallest cross-validation value. For the GAM method, the algorithm is run for eight iterations since preliminary investigations indicate that these are enough for convergence to be reached to sufficient accuracy. As material changes in the choice of h tend to occur in the first few iterations, a cross-validation step is performed in the first four iterations only and then the smoothing parameter is kept fixed. For each combination of density, risk and sample sizes, we simulate 50 sets of data and use all five methods on each to reduce the influence of sampling variability on comparisons between the various methods. The ISE results are summarized in the form of box plots and are shown in Fig. 1. The broken line represents the ISE which would be attained by estimating the risk as constant, equivalent to using an *infinite* smoothing parameter. As we would hope, the methods usually perform substantially better than this base-line.

Overall, the methods do not differ much in their performance, except that the density ratio cross-validation method is much worse than the others for unequal sample sizes. Since the likelihood-based method is intuitively attractive and its performance is at least as good as the other methods, we recommend its use in practice. The GAM method also performs well although it is more computationally demanding; it is therefore our recommended method when additional covariates are available. In Section 5 we compare the kernel regression approach with the GAM method for the Walsall cancer data.

4. Significance

Kelsall and Diggle (1995b) used Monte Carlo sampling to assess the statistical significance of their estimated risk surface. We follow the same basic approach, the main difference being that in a global test of the null hypothesis of no spatial variation in risk we give greater weight to areas with greater numbers of data points.

For each method we obtain an estimate $\hat{\rho}(x)$ which is $\text{logit}\{\hat{p}_h(x)\}$ for the kernel binary regression approach and $\hat{g}(x)$ for the GAM. In each case we adjust $\hat{\rho}(x)$ by removing the average surface level to give $\hat{s}(x)$ with zero mean over the observed data locations,

$$\hat{s}(x) = \hat{\rho}(x) - n^{-1} \sum_{i=1}^{n} \hat{\rho}(x_i).$$

For large data sets the calculation of the mean surface level can be computationally demanding, but the computations can be made more efficient by defining a fine grid of values z_k spanning A such that the x- and y-co-ordinates are both equally spaced by r units, say. If we define N_k to be the number of x_i which fall within $r/2$ of z_k in either the x- or y-co-ordinate, then

$$\hat{s}(x) \simeq \hat{\rho}(x) - n^{-1} \sum_{k} N_k \, \hat{\rho}(z_k).$$

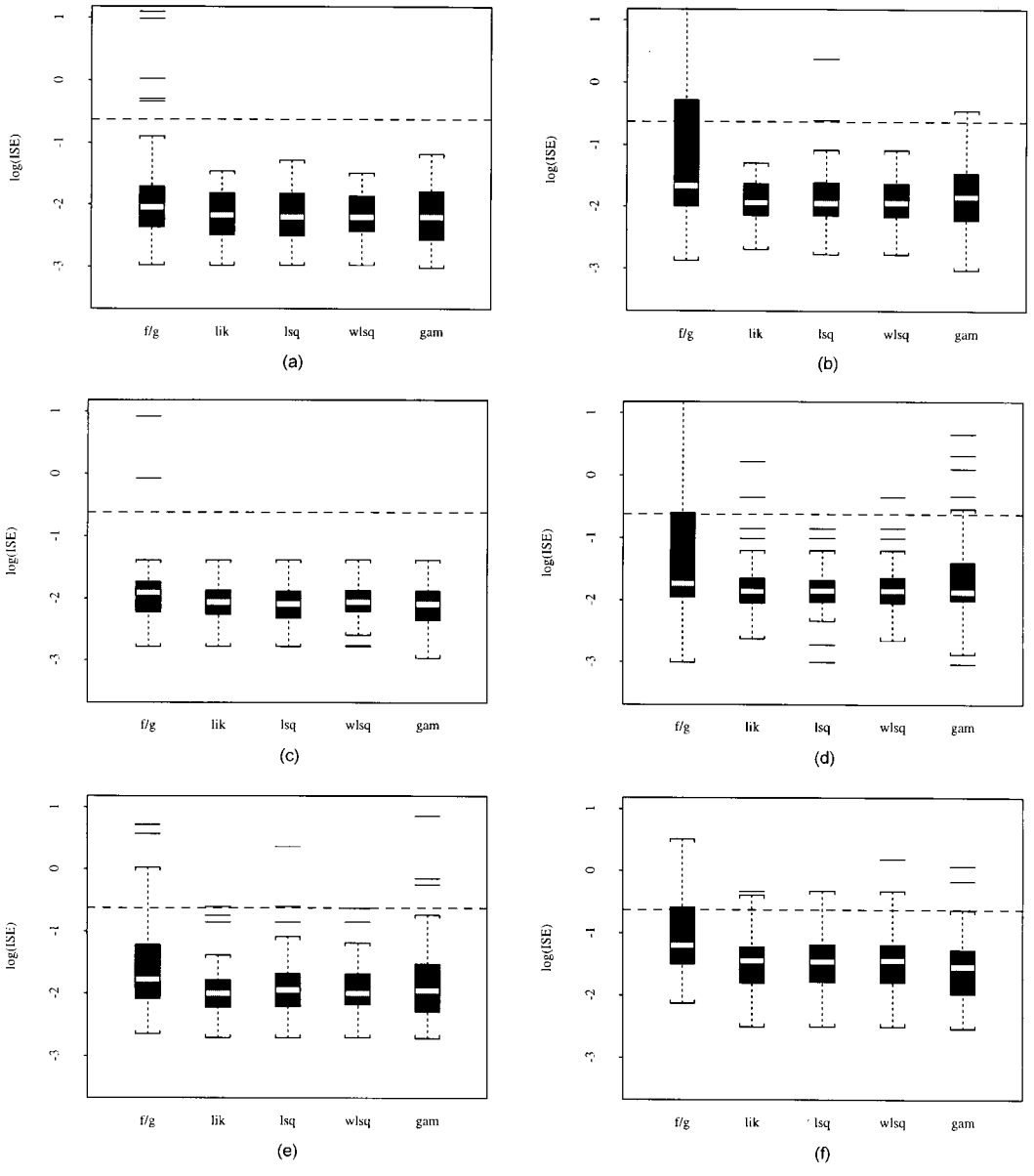

Fig. 1. Box plots to show the log-ISEs of 50 simulations to compare estimation methods (f/g, density ratio cross-validation; lik, likelihood cross-validation; lsq, least squares cross-validation; wlsq, weighted least squares cross-validation; gam, GAM method)—a rapidly fluctuating risk function was used with different combinations of denominator densities and sample sizes n_1 and n_2: (a) density 1, $n_1 = 300$, $n_2 = 300$; (b) density 1, $n_1 = 150$, $n_2 = 600$; (c) density 2, $n_1 = 300$, $n_2 = 300$; (d) density 2, $n_1 = 150$, $n_2 = 600$; (e) density 3, $n_1 = 300$, $n_2 = 300$; (f) density 3, $n_1 = 150$, $n_2 = 600$

We construct tolerance contours which indicate for each x whether $\hat{s}(x)$ is consistent with the null hypothesis H_0: $\rho(x) = c$. We do this by generating new data that are consistent with H_0 but otherwise are similar in distribution to the original data, and we construct a new estimate $\hat{s}_1(x)$. We repeat this m times and construct a p-value surface which, for each x, gives the proportion of values $\hat{s}_i(x)$, $i = 1, \ldots, m$, which are less than the original estimate $\hat{s}_0(x)$. The 0.025 and 0.975 contours of this surface can be added to a grey scale map of $\hat{s}_0(x)$ as 95% tolerance contours to indicate locations of unusually high or low risk.

We can also perform a Monte Carlo test (Barnard, 1963) for overall departure from hypothesis H_0 by using the statistics

$$t_j = n^{-1} \sum_{i=1}^{n} \hat{s}_j(x_i)^2$$

and calculating a p-value as $p = (k + 1)/(m + 1)$ where k is the number of $t_j > t_0$.

The generation of data under hypothesis H_0 depends on the method of estimation. For the kernel regression approach, conditional on the locations of the combined cases and controls, the probability that a given event represents a case will not depend on the spatial location. So to generate data under H_0 we combine the locations of the cases and controls, and randomly label n_1 of them as cases, and the remainder as controls.

When there are covariates we use the GAM approach and assume a model $\text{logit}\{p(x, u)\} = \beta_0 + \beta_1 u + g(x)$, say. Our procedure for generating data that are consistent with H_0: $g(x) = 0$ starts by fitting the reduced model $\text{logit}\{p(x, u)\} = \beta_0 + \beta_1 u$ and calculating fitted probabilities p_i. Conditioning on the numbers of cases and controls and on the locations and covariate values for each individual, we sample n_1 of the $n_1 + n_2$ individuals without replacement according to probabilities proportional to p_i, and we label them as cases. The probability of labelling an individual as a case thus depends on all the covariates except for spatial location.

5. The Walsall cancer data

The Walsall cancer mortality data were introduced in Section 1. We now use our methodology to investigate whether there is any spatial variation in risk of cancers of the lung, stomach and pancreas of which there were 2015, 643 and 318 cases respectively. These *case* data are supplemented by the locations of around three times as many *controls*, taken from the population register of June 1994. There are two possible approaches for selecting controls. The first is to obtain controls by stratified random sampling, such that they have approximately the same joint age and sex distribution as the cases; we then use the simpler kernel regression method of estimation because the effects of covariates are accounted for in the selection of controls. The second approach is to obtain a simple random sample of controls and to take account of the covariates by modelling their effects within the GAM estimation method. For each cancer we obtain stratified and random control groups of the same size. These control group sizes are 5839, 2712 and 1083 for lung, stomach and pancreas cancers respectively.

5.1. Kernel regression

The first step in kernel regression is to choose an appropriate smoothing parameter. We can theoretically derive the value of the likelihood cross-validation criterion (2.1) for infinite h:

$$\mathrm{CV}_2(\infty) = (n_1 + n_2 - 1)(n_1 - 1)^{-p}(n_2 - 1)^{-q},$$

where $p = n_1/(n_1 + n_2)$ and $q = 1 - p$. For ease of interpretation, redefine the cross-validation function to be

$$\mathrm{CV}(h) = \mathrm{CV}_2(h)/\mathrm{CV}_2(\infty),$$

since a value of h for which $\mathrm{CV}(h) < 1$ is then identified as a value which produces a 'better' estimate of relative risk than the null hypothesis of constant risk (equivalent to using an infinite smoothing parameter).

The likelihood cross-validation functions for the three sets of Walsall cancer data using the stratified controls are shown in Fig. 2. The resulting minimizing bandwidths are 500 m, 800 m and ∞. The usefulness of the added line $\mathrm{CV}(h) = 1$ becomes clear on the cross-validation function for the pancreas cancer data, since without it we could easily be misled into believing that the minimum occurred at $h = 1100$ m.

The resulting log-risk surfaces for the lung and stomach cancers are shown in Figs 3 and 4 respectively; in each case, a global test of constant risk based on 500 simulations gave a p-value of 0.002. We also show a risk surface for the pancreas cancer data in Fig. 5 using the value $h = 1100$ m since this corresponded to a local minimum of $\mathrm{CV}(h)$, and also to confirm that the spatial variation in risk is not statistically significant ($p = 0.47$), consistent with the cross-validated h-value of ∞. Note that we use base 2 logarithms so that an increase of 1 in the log-risk surface from one location to another indicates a doubling of risk. This enables an easier interpretation of the maps.

5.2. Generalized additive model with covariates

We now describe an analysis of the Walsall lung cancer data using the GAM method with random controls and taking account of the age and sex covariates in the model. Assume the model

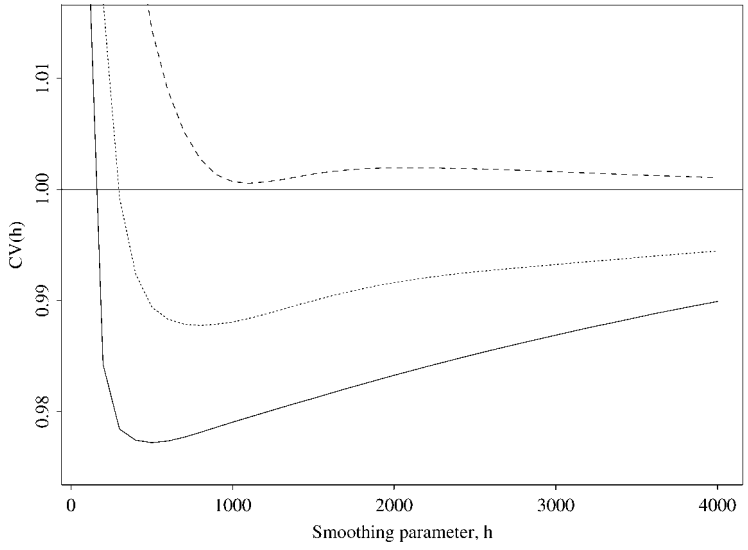

Fig. 2. Likelihood-based cross-validation functions for the Walsall District Health Authority lung (———), stomach (·····) and pancreas (- - - -) cancer mortality data (the smoothing parameter units are metres)

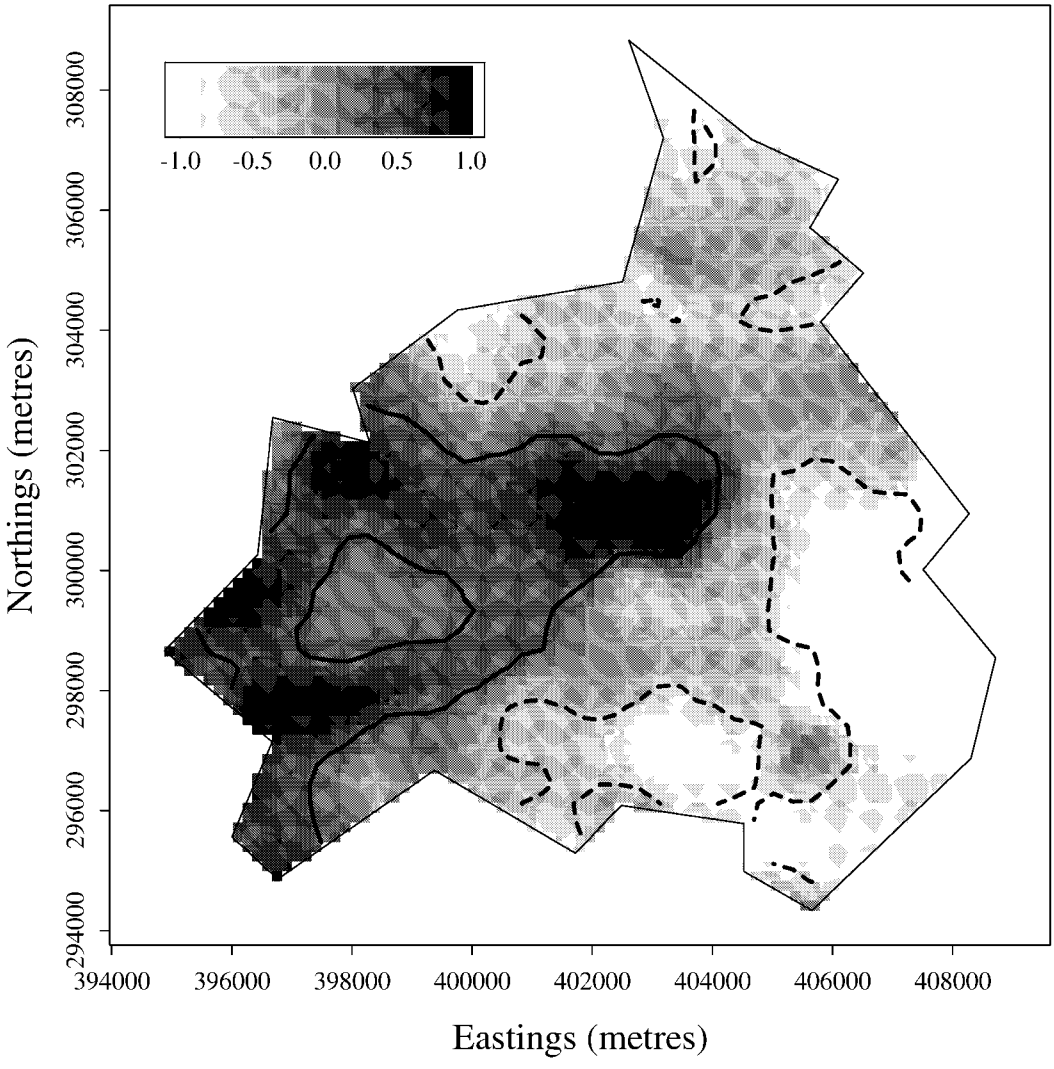

Fig. 3. Logarithmic (base 2) estimated risk surface for lung cancer in the Walsall District Health Authority, with approximate 95% tolerance contours (the test of non-constant risk gave a p-value of 0.002 (based on 500 simulations), and the value of the smoothing parameter was $h = 500$ (the units are metres); kernel regression with a stratified control sample was used): ———, 97.5% contours of the p-value surface; - - - -, 2.5% contours

$$\text{logit}\{p(x, a, s)\} = \beta_0 + \beta_1 s + \beta_2 a + \beta_3 a^2 + g(x), \tag{5.1}$$

where a represents age and s represents sex, $s = 1$ and $s = 0$ denoting males and females respectively. The quadratic effect of age was assumed after simple exploratory analysis. When fitting the model, at each iteration we chose from 15 values of h equally spaced on a log-scale between 100 m and 4000 m. Initial β-parameter estimates were obtained by first fitting a GLM without the $g(x)$ term. GAM iterations were then continued until the choice of smoothing parameter settled at one value, and two further iterations were then conducted. Fig. 6 shows the resulting risk surface estimate for the lung cancer data with 95% tolerance

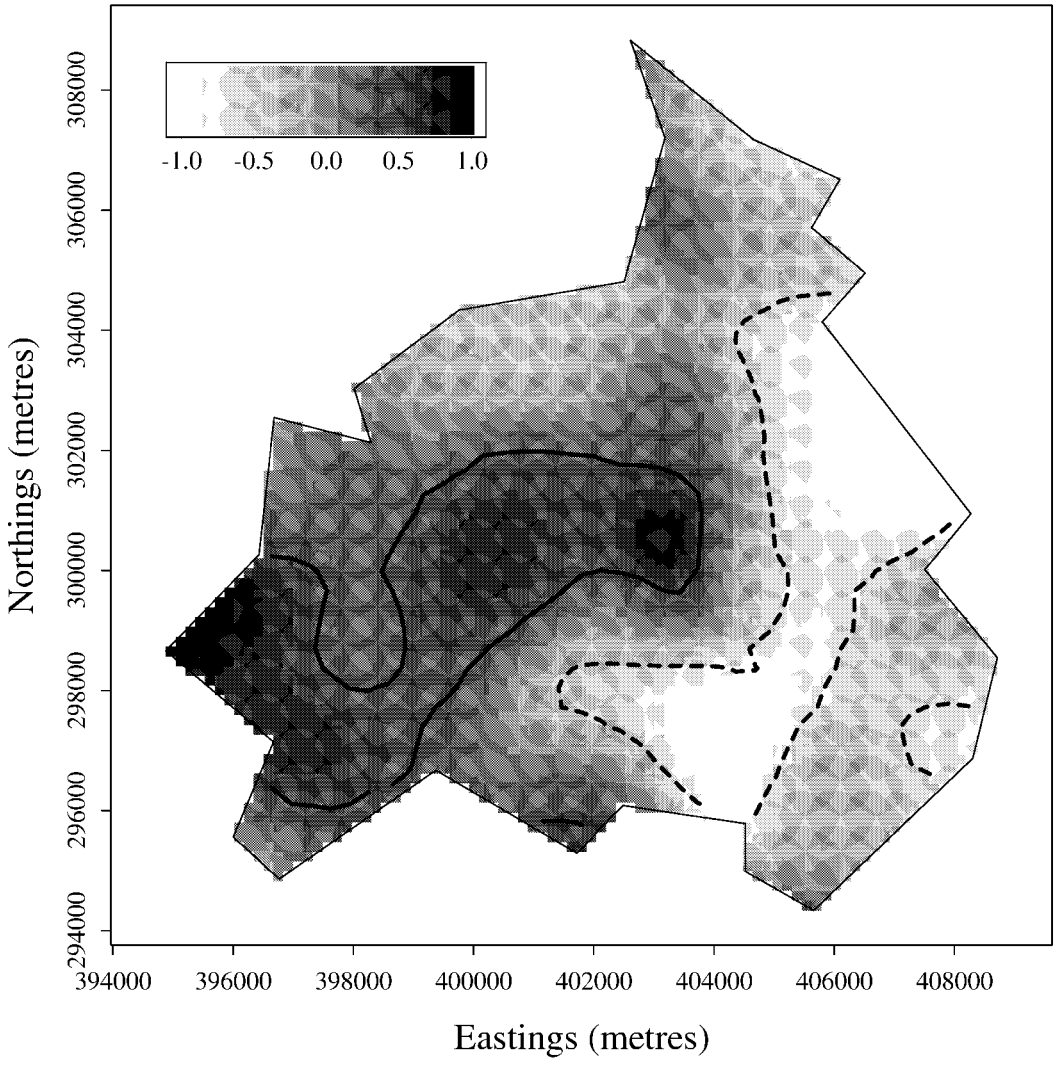

Fig. 4. Logarithmic (base 2) estimated risk surface for stomach cancer in the Walsall District Health Authority, with approximate 95% tolerance contours (the test of non-constant risk gave a p-value of 0.002 (based on 500 simulations), and the value of the smoothing parameter was $h = 800$ (the units are metres); kernel regression with a stratified control sample was used): ———, 97.5% contours of the p-value surface; - - - -, 2.5% contours

contours. The parameter estimates for the age and sex covariates were $\hat{\beta}_1 = 1.36$ (standard error 0.06), $\hat{\beta}_2 = 0.51$ (standard error 0.03) and $\hat{\beta}_3 = -0.0034$ (standard error 0.0002), corresponding to greater risk for males compared with that for women, and increasing up to age 75 years. As expected, the risk surface estimate is very similar to that in Fig. 3 obtained using the stratified control sample approach and kernel regression.

6. Discussion

We have demonstrated that the density ratio and binary kernel regression approaches give

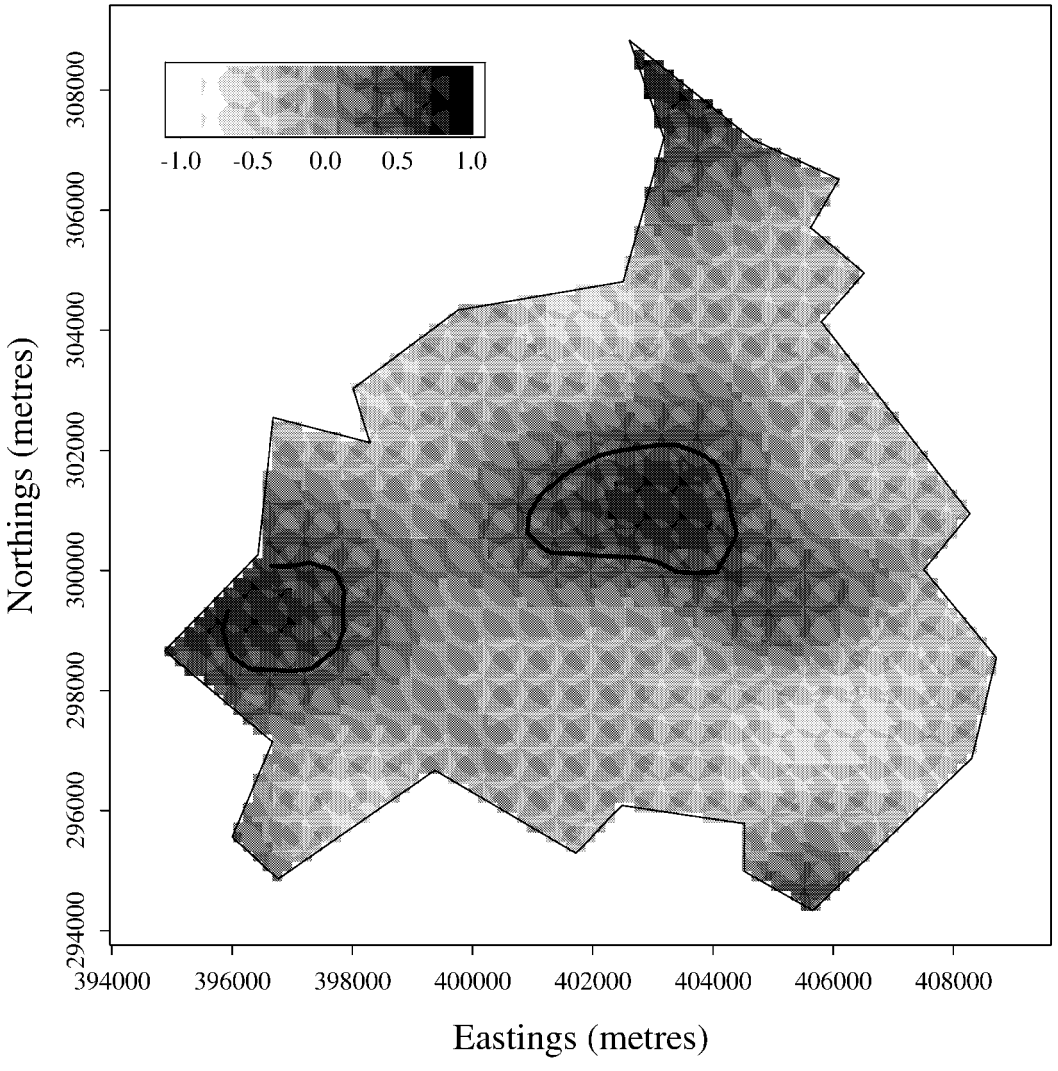

Fig. 5. Logarithmic (base 2) estimated risk surface for pancreas cancer in the Walsall District Health Authority, with approximate 95% tolerance contours (the test of non-constant risk gave a *p*-value of 0.471 (based on 500 simulations), and the value of the smoothing parameter was $h = 1100$ (the units are metres); kernel regression with a stratified control sample was used): ———, 97.5% contours of the *p*-value surface; - - - -, 2.5% contours

the same risk surface estimate, but that the kernel regression method leads to a better criterion for choosing the bandwidth in terms of minimizing the ISE of the final estimate. The GAM is a different estimation procedure that allows explicit covariate adjustments, whereas for the other methods the only way of dealing with covariates is by using stratified controls. Constructing a stratified sample of controls can be much more difficult than obtaining a random sample, particularly when the number of covariates is large. The GAM method is substantially more computer intensive, however, which will mean that the simpler kernel regression method will sometimes be preferable.

The cross-validation criteria proposed for choosing smoothing parameter values are

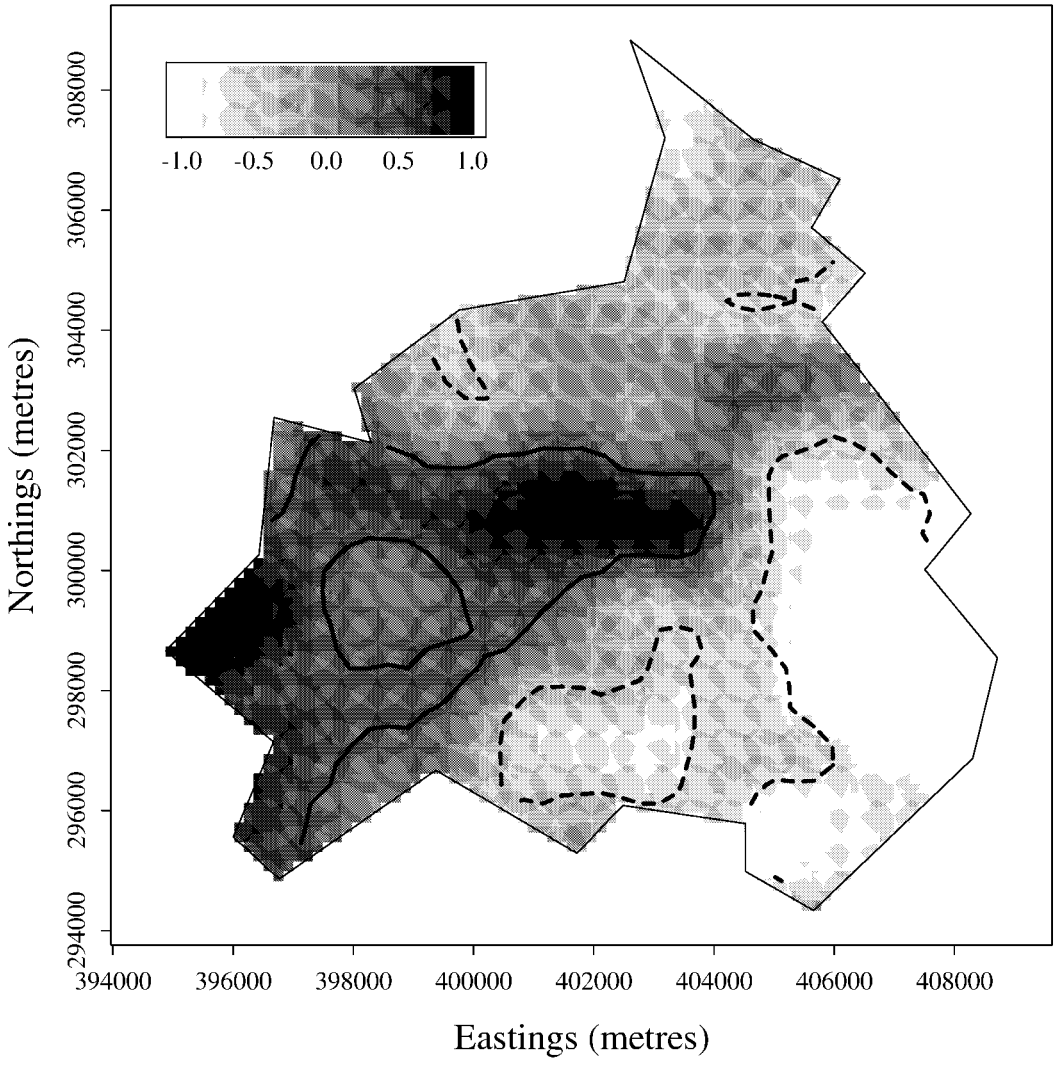

Fig. 6. Logarithmic (base 2) estimated risk surface for lung cancer in the Walsall District Health Authority, with approximate 95% tolerance contours (the test of non-constant risk gave a p-value of 0.005 (based on 200 simulations), and the value of the smoothing parameter was $h = 632$ (the units are metres); the GAM approach with random controls was used, taking account of covariates in the parametric part of the model): ———, 97.5% contours of the p-value surface; - - - -, 2.5% contours

intended as guides only. Ideally, plots of the cross-validation curves should be examined to identify suitable values and to highlight local minima. This strategy was important in the pancreas cancer application. If there is interest in local fluctuations in risk at a particular scale, then a bandwidth value should be used that corresponds to that scale.

In the application of our methods, we found a highly significant variation in risk for lung cancer and stomach cancer, with increased risk in the west and central areas of the region. For pancreas cancers, although not significant, the same pattern of increased risk was observed. By comparing these estimated risk surfaces with a map of social deprivation in Walsall, we see a reasonably close correspondence. It is well known that lung cancer and

stomach cancer have higher prevalence in areas of high social deprivation (e.g. Elliott *et al.* (1992)), probably as a by-product of association between social deprivation and known risk factors such as smoking and eating habits. Our findings are therefore not surprising, but they highlight the ability of our methodology to detect genuine spatial trends in risk. If a measure of deprivation over the region had been available, then it would have been sensible also to include this as a covariate in the GAM to investigate whether there was any residual spatial variation in cancer risk.

The origin of the controls should also be considered when interpreting the estimated risk surfaces. For example, in the Walsall cancer application we assumed that the controls sampled in June 1994 represent the population at risk from which the cancers of 1982–1992 arose. Substantial changes in population structure over this period could lead to areas of estimated high risk which in fact represent only shifts of the local population within the region.

Acknowledgement

We thank Ruth Wain (Walsall Health, Walsall) for providing the Walsall cancer data.

References

Azzalini, A., Bowman, A. W. and Härdle, W. (1989) On the use of nonparametric regression for model checking. *Biometrika*, **76**, 1–11.

Barnard, G. A. (1963) Discussion on The spectral analysis of point processes (by M. S. Bartlett). *J. R. Statist. Soc.* B, **25**, 294.

Barry, J., Crowder, M. and Diggle, P. (1997) Parametric estimation of the variogram. *Technical Report ST-97-06*. Department of Mathematics and Statistics, Lancaster University, Lancaster.

Bithell, J. F. (1990) An application of density estimation to geographical epidemiology. *Statist. Med.*, **9**, 691–701.

———(1992) Statistical methods for analysing point-source exposures. In *Geographical and Environmental Epidemiology: Methods for Small Area Studies* (eds P. Elliott, J. Cuzick, D. English and R. Stern), pp. 221–230. Oxford: Oxford University Press.

Copas, J. B. (1983) Plotting p against x. *Appl. Statist.*, **32**, 25–31.

Elliott, P., Hills, M., Beresford, J., Kleinschmidt, I., Jolley, D., Pattenden, S., Rodrigues, L., Westlake, A. and Rose, G. (1992) Incidence of cancer of the larynx and lung near incinerators of waste solvents and oils in Great Britain. *Lancet*, **339**, 854–858.

Fan, J. Q., Heckman, N. E. and Wand, M. P. (1995) Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Am. Statist. Ass.*, **90**, 141–150.

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

Härdle, W. (1990) *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

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

Kelsall, J. E. (1996) Kernel smoothing for application in environmental epidemiology. *PhD Thesis*. Lancaster University, Lancaster.

Kelsall, J. E. and Diggle, P. J. (1995a) Kernel estimation of relative risk. *Bernoulli*, **1**, 3–16.

———(1995b) Non-parametric estimation of spatial variation in relative risk. *Statist. Med.*, **14**, 2335–2342.

Lawson, A. B. and Williams, F. L. R. (1993) Applications of extraction mapping in environmental epidemiology. *Statist. Med.*, **12**, 1249–1258.

McCullagh, P. (1983) Quasi-likelihood functions. *Ann. Statist.*, **11**, 59–67.

McCullagh, P. and Nelder J. A. (1989) *Generalized Linear Models*. London: Chapman and Hall.

Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*. London: Chapman and Hall.