Biostatistics (2006), **7**, 3, *pp.* 438–455
doi:10.1093/biostatistics/kxj017
Advance Access publication on January 20, 2006

Health-exposure modeling and the ecological fallacy

JON WAKEFIELD*

Departments of Statistics and Biostatistics, University of Washington, Seattle, WA, USA
jonno@u.washington.edu

GAVIN SHADDICK

Department of Mathematical Sciences, University of Bath, Bath, UK

SUMMARY

Recently, there has been an increased interest in modeling the association between aggregate disease counts and environmental exposures measured, for example via air pollution monitors, at point locations. This paper has two aims: first, we develop a model for such data in order to avoid ecological bias; second, we illustrate that modeling the exposure surface and estimating exposures may lead to bias in estimation of health effects. Design issues are also briefly considered, in particular the loss of information in moving from individual to ecological data, and the at-risk populations to consider in relation to the pollution monitor locations. The approach is investigated initially through simulations, and is then applied to a study of the association between mortality in those over 65 in the year 2000 and the previous year's SO_2, in London. We conclude that the use of the proposed model can provide valid inference, but the use of estimated exposures should be carried out with great caution.

Keywords: Ecological fallacy; Environmental epidemiology; Exposure modeling; Quasi-likelihood; Spatial epidemiology.

1. INTRODUCTION

Recently, a great deal of attention has been paid to the investigation of associations between health outcomes and environmental exposures that may be measured in air, water, or soil. Population and health data are often routinely available in ecological, that is group, form while the exposure data typically consist of a set of values recorded at monitor sites or via one-off sampling. The exposure information is usually spatially sparse, which has recently lead to the modeling of an exposure surface. We primarily consider models appropriate for point sampling of environmental rather than behavioral exposures such as dietary, smoking, and alcohol variables; information on behavioral variables is obtained from individuals at specific residential locations, often via surveys. The model we introduce in Section 3 may be used for behavioral exposures, but the estimation of an exposure surface would not be of interest since behavioral variables do not generally exhibit spatial structure. Exposures that are conducive to examination via ecological designs and which are amenable to analysis with the model developed in this paper

*To whom correspondence should be addressed.

include chronic air pollution (for examples, see Table 7.2 of Pope and Dockery, 1996), water constituents (e.g. Maheswaran *et al.*, 1999), and soil contaminants (e.g. Elliott *et al.*, 2000).

A number of authors have considered the modeling of exposure surfaces. Le *et al.* (1996) develop theory based on a multivariate normal distribution to model air pollution variables, Gelfand *et al.* (2001) use a Gaussian random field (GRF) model for the modeling of ozone, Tonellato (2001) considers the modeling of carbon monoxide at multiple sites, and Shaddick and Wakefield (2002) model the spatiotemporal variability of four pollutants in London. More recently, a number of authors have combined health data with modeled exposures. For example, Zidek *et al.* (1998) extend the work of Le *et al.* (1996) to examine the association between daily hospital admissions for respiratory disease and sulfate concentrations; Carlin *et al.* (1999) examine the relationship between pediatric asthma emergency room visits and ozone, where the latter are modeled using kriging within a geographic information systems (GIS); and Zhu *et al.* (2003) extend this analysis by assuming the ozone measures arise from a continuous, stationary spatial process whose parameters are estimated using Bayesian methods. The above authors do not consider correcting for ecological bias; assuming that associations observed at the level of the area hold for the individuals within the areas can lead to the so-called ecological fallacy (Selvin, 1958). Ecological bias can manifest itself in a variety of ways; the one that we concentrate on is 'pure specification bias', which arises under aggregation of a non-linear model. The aims of this paper are two-fold. First, we develop a convolution model that avoids pure specification bias due to the use of an incorrect mean function. Second, we illustrate the problems of the use of estimated exposures within a health model.

As an example of a study for which the methods of this paper are intended, we examine the association between respiratory mortality in the year 2000 in those over 65 in inner London and the previous year's SO_2, measured in parts per billion (ppb). The latter is available as the yearly average of (daily) values at each of the 16 monitor sites, and is a concentration. A major problem with such studies is that the density of exposure monitors is insufficient to fully characterize the exposure surface for a complete geographical study region. To illustrate, population and health data were extracted for all enumeration districts (EDs) whose centroids lie within 1 km of at least one of the monitor sites (an ED is a census-defined geographical area that contains on average 400 individuals); 1 km was chosen as this radius is sufficiently large to show the exposure characterization problems.

Table 1 reports summary statistics for the study; the populations are not integers since they have been adjusted for undercount and migration (Simpson *et al.*, 1996). Figure 1 shows the locations of the 16 monitor sites. A plot of mortality risk versus SO_2 (at the ecological level) indicates no clear association. We emphasize that in this application we have observed mortality and population information at each of the 1027 EDs whose centroids lie within 1 km of a monitor, but exposure is only measured at the 16 monitors.

The structure of this paper is as follows: In Section 2 we indicate a number of inadequacies with previous approaches, and in Section 3 suggest a new model. In Section 4 we demonstrate the use of this model on simulated data, and in Section 5 return to the motivating example. Section 6 provides a concluding discussion.

2. Previous approaches

Consider a study region A consisting of K sub-areas, A_k, for which population data, N_k, and disease data, Y_k, are available, $k = 1, \ldots, K$. We assume a univariate exposure and no confounders. Exposure data x_m are available from a set of pollution monitors within the study region, at locations $\boldsymbol{s}_m$, $m = 1, \ldots, M$. A naive ecological model is given by

$$Y_k|\overline{x}_k, \boldsymbol{\beta}^\star \sim_{\text{ind}} \text{Poisson}\{N_k \exp(\beta_0^\star + \beta_1^\star \overline{x}_k)\}, \tag{2.1}$$

where $\boldsymbol{\beta}^\star = (\beta_0^\star, \beta_1^\star)$ and $\overline{x}_k$ is the observed mean exposure within area k, $k = 1, \ldots, K$.

Table 1. *Summary statistics for the respiratory mortality study. The second column identifies the monitor with the label on Figure 1*

Monitor site	Label	Number of EDs	Number of cases	Population (over 65)	Incidence rate $\times 10^3$	Yearly mean SO_2 (ppb)
Bromley	1	33	18	2365.0	7.61	0.90
Bexley	2	28	13	1847.2	7.04	2.83
Bloomsbury	3	84	20	3377.6	5.92	4.96
Brent	4	31	12	1494.5	8.03	1.64
London Bridge Place	5	102	46	5218.9	8.81	2.64
Cromwell Road	6	124	16	3808.0	4.20	4.38
Eltham	7	16	16	1615.9	9.90	2.06
Hillingdon	8	12	4	929.1	4.31	3.80
Lewisham	9	42	53	2688.7	19.7	2.94
Marylebone Road	10	100	16	4138.0	3.87	5.03
North Kensington	11	99	40	4332.5	9.23	2.44
Southwark	12	87	43	5497.4	7.82	2.88
Teddington	13	23	24	1763.4	13.6	2.21
Southwark roadside	14	57	35	3015.8	11.6	3.66
Sutton	15	42	43	2912.2	14.8	3.14
West London	16	147	28	4260.4	6.57	0.31
Totals/means		1027	427	49 264.6	9.94	2.86

Fig. 1. Locations of the 16 pollution monitor sites in London, each circle is of radius 1 km. Health and population data are from all EDs whose centroids lie within 1 km of any monitor. The names of the monitor sites are given in Table 1. The River Thames is marked, and the light lines denote boundaries of London boroughs.

To illustrate the problems with (2.1), consider individual i in area k, let Y_{ki} denote a Bernoulli disease indicator, and assume the individual-level model is

$$Y_{ki}|x_{ki}, \boldsymbol{\beta} \sim_{\text{ind}} \text{Bernoulli}\{p(x_{ki}, \boldsymbol{\beta})\}, \tag{2.2}$$

for $k = 1, \ldots, K$, $i = 1, \ldots, N_k$. For a rare disease, assume $p(x, \boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x)$ so that $\boldsymbol{\beta} = (\beta_0, \beta_1)$. We emphasize that the individual-level parameter of interest is β_1, while in (2.1), $\beta_1^\star$ is the ecological association, and ecological bias will in general result in $\beta_1 \neq \beta_1^\star$. The characterization of ecological bias has seen a great deal of attention (see, for example Richardson *et al.*, 1987; Piantadosi *et al.*, 1988; Greenland and Morgenstern, 1989; Greenland, 1992; Greenland and Robins, 1994; Diggle and Elliott, 1995; Plummer and Clayton, 1996; Wakefield and Salway, 2001; Wakefield, 2003, 2004).

In the aggregate setting, we do not know the individual responses, Y_{ki}, but rather the sum Y_k. Letting $\boldsymbol{x}_k = \left(x_{k1}, \ldots, x_{kN_k}\right)^{\text{T}}$ denote the collection of the exposures for the individuals of area k, we have

$$\text{E}[Y_k|\boldsymbol{x}_k, \boldsymbol{\beta}] = N_k q_k,$$

where

$$q_k = \frac{\text{e}^{\beta_0}}{N_k} \sum_{i=1}^{N_k} \exp(\beta_1 x_{ki}) \tag{2.3}$$

is the average risk of the individuals in area k. If we know the collection $\boldsymbol{x}_k$ but not the linkage with individuals, then each of the N_k responses are Bernoulli with probability (2.3), but they are not independent (since we are sampling without replacement from $\boldsymbol{x}_k$) and so Y_k is not binomial with parameters N_k and q_k; we derive the appropriate likelihood in Section 3. An alternative approach (related to block Kriging) is to model the exposure surface, $x(\boldsymbol{s})$, for $\boldsymbol{s} \in A$, form the average

$$\overline{x}_k = \int_{A_k} x(\boldsymbol{s})\, f_k(\boldsymbol{s}) \text{d}\boldsymbol{s}, \tag{2.4}$$

where $f_k(\boldsymbol{s})$ represents the population density in area k at location $\boldsymbol{s}$, and then substitute this mean into (2.1). Such an approach leads to ecological bias since the risk function is evaluated at the mean exposure, while (2.3) shows that we should calculate the mean of the risk functions. Zhu *et al.* (2003), building on Gelfand *et al.* (2001), use such an approach within a Bayesian hierarchical model.

Prentice and Sheppard (1995) propose an 'aggregate data' method in which exposures x_{kj}, $j = 1, \ldots, m_k \leqslant N_k$, are available on a subset of individuals (for further details of this approach, see Sheppard and Prentice, 1995 and Guthrie and Sheppard, 2001). An estimate of (2.3) is given by

$$\widehat{q}_k = \frac{\text{e}^{\beta_0}}{m_k} \sum_{j=1}^{m_k} \exp(\beta_1 x_{kj}) \tag{2.5}$$

which, together with the variance of Y_k, allows an estimating equation for $\boldsymbol{\beta}$ to be constructed. If $m_k < N_k$, then the estimating equation is biased, but Prentice and Sheppard (1995) obtain an expression for this bias, and use this to provide an unbiased estimating equation.

A second approach (Richardson *et al.*, 1987) assumes that $p_k(\cdot|\boldsymbol{\phi}_k)$ is the distribution of exposure in area k, with parameters $\boldsymbol{\phi}_k$, in which case the average risk is

$$\text{e}^{\beta_0} \int_{x(\boldsymbol{s}):\boldsymbol{s}\in A_k} \exp(\beta_1 x)\, p_k(x|\boldsymbol{\phi}_k) \text{d}x, \tag{2.6}$$

and the link with (2.3) is revealed if we replace $f_k(\boldsymbol{x}|\boldsymbol{\phi}_k)$ by a discrete distribution on $x_{k1}, \ldots, x_{kN_k}$. Pure specification bias occurs with the use of (2.1) because, unless the exposure is constant within A_k, integrating a non-linear risk model leads to the model changing form. As an illustration, for a normal

within-area distribution, $x_{ki} \sim_{\text{iid}} N(\overline{x}_k, s_k^2)$, so that $\boldsymbol{\phi}_k = (\overline{x}_k, s_k^2)$, (2.6) takes the form

$$\exp(\beta_0 + \beta_1 \overline{x}_k + \beta_1^2 s_k^2/2). \tag{2.7}$$

A plausible model that is amenable to analytic study (Wakefield, 2003) is to assume that $s_k^2 = a + b\overline{x}_k$ so that if $b > 0$ the variance increases with the mean, a behavior that is often observed with environmental exposures (e.g. Ott, 1994). This choice leads to (2.7) taking the ecological form

$$\exp(\beta_0 + a\beta_1^2/2 + \beta_1 \overline{x}_k + b\beta_1^2 \overline{x}_k/2), \tag{2.8}$$

so that, in terms of the naive ecological model (2.1), we have

$$\beta_0^\star = \beta_0 + a\beta_1^2/2, \quad \beta_1^\star = \beta_1 + b\beta_1^2/2, \tag{2.9}$$

illustrating that bias will result unless $b = 0$, that is unless the variance is independent of the mean. It is clear in this case (and true more generally) that pure specification bias is small if β_1, and/or the within-area variabilities in exposure, are close to zero. Hence, for example, we may conclude that in the study of Zhu *et al.* (2003), pure specification bias will be small since the study areas are zip codes and the main exposure contrasts are temporal rather than spatial. In such a context, modeling an exposure surface is likely to give small benefits, however, and may even be detrimental, as we show in Section 4.2.

3. A CONVOLUTION MODEL

3.1 *Model development*

The likelihood, under the assumptions of Section 2 and when all individual-level exposures are available, is the convolution

$$\Pr(Y_k = y_k | \boldsymbol{x}_k) = \sum_{\boldsymbol{y}_k \in C_{y_k}} \prod_{i=1}^{N_k} \Pr(Y_{ki} = y_{ki} | x_{ki}) = \sum_{\boldsymbol{y}_k \in C_{y_k}} \prod_{i=1}^{N_k} p_{ki}^{y_{ki}} (1 - p_{ki})^{1 - y_{ki}}, \tag{3.1}$$

where $\boldsymbol{y}_k = \left(y_{k1}, \ldots, y_{kN_k}\right)^{\mathrm{T}}$, $p_{ki} = p(x_{ki})$ is the risk model evaluated at x_{ki}, and C_{y_k} is the set containing the $\binom{N_k}{y_k}$ ways of assigning Y_k cases to N_k individuals. In general, (3.1) will be computationally expensive to enumerate (since N_k is typically large), but in the case of a rare event, each of the Bernoulli random variables may be approximated by a Poisson random variable, and with the log-linear risk model $p_{ki} = \mathrm{e}^{\beta_0 + \beta_1 x_{ki}}$, we obtain the convolution model

$$Y_k | \boldsymbol{x}_k \sim_{\text{ind}} \text{Poisson}\left\{ \mathrm{e}^{\beta_0} \sum_{i=1}^{N_k} \exp(\beta_1 x_{ki}) \right\}. \tag{3.2}$$

This distribution should still be viewed as group level because we have individual-level exposures, but only aggregate disease counts and there is no linkage between individual-level outcomes and exposures. However, the use of this model removes pure specification bias.

Usually, the full exposure information $x_{ki}, i = 1, \ldots, N_k$, will be unavailable. Suppose, however, that m_k exposures, x_{kj}, are measured at locations $\boldsymbol{s}_{kj}$, $j = 1, \ldots, m_k$. One possible use of this information is to allocate N_{kj} individuals to measurement x_{kj}. For example, suppose we have populations N_{kj} within ED j contained within region k, and exposures, x_{kj} at ED centroids, $\boldsymbol{s}_{kj}$, but disease counts, Y_k, at a coarser geographical scale (for example the monitor regions in the motivating example), we may then allocate ED

population N_{kj} to exposure x_{kj}. One may then replace (3.2) with

$$Y_k|\boldsymbol{x}_k \sim_{\text{ind}} \text{Poisson}\left\{ \mathrm{e}^{\beta_0} \sum_{j=1}^{m_k} N_{kj} \exp(\beta_1 x_{kj}) \right\}. \tag{3.3}$$

If we take $N_{kj} = N_k/m_k$, so that we divide the population equally, then

$$Y_k|\boldsymbol{x}_k \sim_{\text{ind}} \text{Poisson}\left\{ N_k \frac{\mathrm{e}^{\beta_0}}{m_k} \sum_{j=1}^{m_k} \exp(\beta_1 x_{kj}) \right\}. \tag{3.4}$$

Comparison with (2.5) reveals we have a parametric version of the aggregate method of Prentice and Sheppard (1995). If $m_k < N_{kj}$, then this model is susceptible to ecological bias, and explains the finite-correction bias suggested for the aggregate method. The key to minimizing ecological bias is to have a fine enough partition of space at which exposure measurements are available, relative to the spatial exposure variability.

3.2 *Inference with known exposures*

Inference for the convolution model (3.3), with the exposures x_{kj} known, is easily carried out via likelihood, with the extension to quasi-likelihood being immediate. The Poisson log-likelihood corresponding to (3.3) is not a generalized linear model since we do not have a linear predictor, but may be maximized with respect to β_0 in closed form to give the profile log-likelihood for β_1:

$$l_p(\beta_1) = -y_+ \log\left(\sum_{k=1}^{K} \sum_{j=1}^{m_k} N_{kj} \exp\{\beta_1 x_{kj}\} \right) + \sum_{k=1}^{K} y_k \log\left(\sum_{j=1}^{m_k} N_{kj} \exp\{\beta_1 x_{kj}\} \right),$$

which is straightforward to maximize.

In most ecological studies, the sample sizes are large and asymptotic inference is likely to be accurate, at least for simple models. For the convolution model (3.3), the expected information is given by

$$\boldsymbol{I}^{\mathrm{C}}(\boldsymbol{\beta}) = \begin{bmatrix} \sum_{k=1}^{K} \sum_{j=1}^{m_k} N_{kj}\, p_{kj} & \sum_{k=1}^{K} \sum_{j=1}^{m_k} N_{kj} x_{ki}\, p_{kj} \\ \sum_{k=1}^{K} \sum_{j=1}^{m_k} N_{kj} x_{ki}\, p_{kj} & \sum_{k=1}^{K} \left(\sum_{j=1}^{m_k} N_{kj} x_{kj}\, p_{kj} \right)^2 \Big/ \sum_{j=1}^{m_k} N_{kj}\, p_{kj} \end{bmatrix}, \tag{3.5}$$

where $p_{kj} = \exp(\beta_0 + \beta_1 x_{kj})$. Quasi-likelihood is based on assuming that $\text{var}(Y_k|\boldsymbol{\beta}) = \kappa \times \text{E}[Y_k|\boldsymbol{\beta}]$ with $\text{cov}(Y_k, Y_{k'}) = 0$ for $k \neq k'$. Point estimates are the same as under maximum likelihood, and standard errors are multiplied by $\sqrt{\widehat{\kappa}}$, using the method-of-moments estimator

$$\widehat{\kappa} = \frac{1}{K-2} \sum_{k=1}^{K} \frac{(y_k - \widehat{\mu}_k)^2}{\widehat{\mu}_k},$$

with $\widehat{\mu}_k = \sum_{j=1}^{m_k} N_{kj} \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_{kj})$ (McCullagh and Nelder, 1989). If the variance is proportional to the mean, and the data are independent, the asymptotic distribution of the estimator for $\boldsymbol{\beta}$, as $K \to \infty$ and with $m_k = N_k$, is given by

$$\boldsymbol{I}^{\mathrm{C}}(\boldsymbol{\beta})^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_{\mathrm{d}} N(\boldsymbol{0}, \kappa \boldsymbol{I}_2).$$

If $m_k < N_k$, there will be ecological bias, as with the uncorrected estimator of Prentice and Sheppard (1995). Quasi-likelihood is appealing since it is straightforward to implement and provides a consistent

estimator so long as the first two moments are correctly specified. If residual spatial dependence is present, then a more complex approach is required; in this case, random-effects models are appealing, and computation via Markov chain Monte Carlo (MCMC) is convenient. Section 6 gives brief details of a model that allows for residual spatial dependence.

3.3 *Information considerations*

The loss of information in moving from individual to aggregate data may be quantified via examination of the respective information matrices. For individual-level Poisson data,

$$Y_{ki}|x_{ki} \sim_{\text{ind}} \text{Poisson}(p_{ki}), \tag{3.6}$$

where $p_{ki} = \exp(\beta_0 + \beta_1 x_{ki})$, and

$$\boldsymbol{I}^{\text{I}}(\boldsymbol{\beta}) = \begin{bmatrix} \sum_{k=1}^{K}\sum_{i=1}^{N_k} p_{ki} & \sum_{k=1}^{K}\sum_{i=1}^{N_k} x_{ki}\, p_{ki} \\ \sum_{k=1}^{K}\sum_{i=1}^{N_k} x_{ki}\, p_{ki} & \sum_{k=1}^{K}\sum_{i=1}^{N_k} x_{ki}^2\, p_{ki} \end{bmatrix}. \tag{3.7}$$

For direct comparison between (3.5) and (3.7), we take $m_k = N_k$ so that $N_{kj} = 1$. In (3.5) the element $\boldsymbol{I}_{22}^{C}$, that represents the information for the parameter of interest β_1, may be written as

$$\sum_{k=1}^{K}\left[\sum_{i=1}^{N_k} x_{ki}^2\, p_{ki} - \left\{\frac{1}{\sum_{i=1}^{N_k} p_{ki}}\left[\left(\sum_{i=1}^{N_k} x_{ki}^2\, p_{ki}\right)\left(\sum_{i=1}^{N_k} p_{ki}\right) - \left(\sum_{i=1}^{N_k} x_{ki}\, p_{ki}\right)^2\right]\right\}\right]$$

so that the term within braces represents the loss of information associated with the convolution; this term is zero if there is no within-area variability in exposure, and increases as the within-area variability increases.

We now present a simple example to illustrate the loss of information in moving from individual to ecological outcomes. Specifically, we examine the asymptotic efficiency in using the convolution model (3.2) relative to the individual model (3.6). We assume there are $N_k = 400$ individuals in each of $K = 1000$ areas, so that we have roughly the same number of areas and the same population sizes as in the motivating study. Within-area exposures are assumed normal with $\bar{x}_k = 2 + 3 \times (k-1)/(K-1)$ and $s_k^2 = b\bar{x}_k$, $k = 1, \ldots, K$. Table 2 reports the efficiencies for a number of values of β_0, β_1, and b. We also give the ratio of the between-area variability in exposure to the sum of the within- and between-area variability; an ecological study is likely to be carried out when this ratio is large since between-area variability in exposure is being exploited. The final column gives the bias of the ecological model, $\text{E}[\widehat{\beta}_1^{\star}] - \beta_1 = b\beta_1^2/2$; we see overestimation in this situation in which the exposure variance increases with the mean. Line 1 of the table has a weak mean–variance relationship and a relative risk close to 1, and hence the bias is small; the variance of the convolution estimator is increased by 47% relative to the individual estimator. In other cases, the variance of the convolution estimator is 2.03–2.66 times greater than the variance of the individual estimator.

3.4 *Inference with estimated exposures*

We now consider the situation in which the exposures, x_{kj}, in model (3.3) are unknown. Estimation of these exposures, based on monitored exposures $\boldsymbol{x}_m$, at locations $\boldsymbol{s}_m$, $m = 1, \ldots, M$, may be carried out if an appropriate exposure model is available to interpolate across the study region. We take a Bayesian

Table 2. *Comparison of information in different designs. 'Between' refers to the between-area variability in exposure means, that is var($\overline{x}_k$), and 'total' the sum of the between- and the average within-area variability* $\mathrm{E}[s_k^2]$*; $\mathrm{var}_\mathrm{C}(\widehat{\beta}_1)$ and $\mathrm{var}_\mathrm{I}(\widehat{\beta}_1)$ are the asymptotic variances of $\widehat{\beta}_1$ under the convolution and individual models*

b	Exposure ratio Between/total	β_0	β_1	$\frac{\mathrm{var}_\mathrm{C}(\widehat{\beta}_1)}{\mathrm{var}_\mathrm{I}(\widehat{\beta}_1)}$	Bias of ecological model
0.1	0.68	−5	log 1.2	1.47	0.002
0.2	0.52	−9	log 2	2.03	0.05
0.2	0.52	−10	log 3	2.27	0.12
0.3	0.42	−9	log 2	2.38	0.07
0.3	0.42	−10	log 3	2.66	0.18

approach to modeling with unknown exposures, which is convenient to reveal the implications of a number of approximations.

Denote by $\boldsymbol{y}^K = (y_1, \ldots, y_K)^\mathrm{T}$ the vector of observed disease counts; $\boldsymbol{x}^K = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K)^\mathrm{T}$, with $\boldsymbol{x}_k = \left(x_{k1}, \ldots, x_{km_k}\right)^\mathrm{T}$, $k = 1, \ldots, K$, the set of unknown exposures; and $\boldsymbol{x}^M = (x_1, \ldots, x_M)^\mathrm{T}$ the set of observed exposures. Adopting a Bayesian approach to inference and exploiting conditional independencies, the joint posterior, over the unknown parameters and exposures, is given by

$$p(\boldsymbol{\beta}, \boldsymbol{x}^K | \boldsymbol{y}^K, \boldsymbol{x}^M) p(\boldsymbol{y}^K, \boldsymbol{x}^M) = p(\boldsymbol{\beta} | \boldsymbol{x}^K, \boldsymbol{y}^K) p(\boldsymbol{x}^K | \boldsymbol{x}^M). \tag{3.8}$$

The posterior for $\boldsymbol{\beta}$ is given by

$$p(\boldsymbol{\beta} | \boldsymbol{x}^K, \boldsymbol{y}^K) \propto \prod_{k=1}^{K} p(y_k | \boldsymbol{\beta}, \boldsymbol{x}_k) p(\boldsymbol{\beta}),$$

where the predictive distribution $p(\boldsymbol{x}^K | \boldsymbol{x}^M)$ may be obtained by assuming a parametric form. We illustrate by assuming a GRF model for the log exposures. Letting $\boldsymbol{\psi}_x$ represent the parameters of this model, the predictive distribution in (3.8) is given by

$$p(\boldsymbol{x}^K | \boldsymbol{x}^M) = \int \prod_{k=1}^{K} p(\boldsymbol{x}_k | \boldsymbol{\psi}_x) p(\boldsymbol{\psi}_x | \boldsymbol{x}^M) \mathrm{d}\boldsymbol{\psi}_x, \tag{3.9}$$

where

$$p(\boldsymbol{\psi}_x | \boldsymbol{x}^M) \propto \prod_{m=1}^{M} p(x_m | \boldsymbol{\psi}_x) p(\boldsymbol{\psi}_x). \tag{3.10}$$

Under a GRF model, each of the distributions $p(\boldsymbol{x} | \boldsymbol{\psi}_x)$ in (3.9) is multivariate lognormal. Since $\boldsymbol{x}^K$ is present in both terms on the right-hand side of (3.8), a fully Bayesian approach would require simultaneous estimation of the health and exposure parameters. This has the advantage of allowing feedback between the health and exposure models, but the disadvantage is that implementation, via MCMC, is computationally expensive since the dimension of the estimated exposure vector $\left(\sum_{k=1}^{K} m_k\right)$ is high. We discuss two approximations that ease this computational burden.

An approximation that cuts the link between the two components of (3.8) (the health and exposure models) takes an estimate $\widehat{\boldsymbol{x}}_k$, and then substitutes this into the likelihood to give $p(\boldsymbol{y}_k | \boldsymbol{\beta}, \widehat{\boldsymbol{x}}_k)$; this allows separate computation of the exposure and health models, but is dangerous since the errors-in-variables

aspect of using an estimated exposure is not acknowledged. A more sophisticated approach that still allows separate computation but acknowledges the uncertainty in $\widehat{\boldsymbol{x}}_k$ is to approximate the predictive distribution. More precisely, we may assume the three-stage model:

Stage 1, Health model.

$$Y_k|\boldsymbol{x}_k, \boldsymbol{\beta} \sim_{\text{ind}} \text{Poisson}\left\{ \mathrm{e}^{\beta_0} \sum_{i=1}^{N_k} \exp(\beta_1 x_{ki}) \right\}, \quad k = 1, \ldots, K.$$

Stage 2, Exposure model. Let $z_{kj} = \log x_{kj}$ and $\boldsymbol{z}^K = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K)^{\mathrm{T}}$, with $\boldsymbol{z}_k = \left(z_{k1}, \ldots, z_{km_k}\right)^{\mathrm{T}}$, $k = 1, \ldots, K$. If we ignore the uncertainty in the posterior for $\boldsymbol{\psi}_x$

$$p(\boldsymbol{x}_k|\boldsymbol{x}^M) \approx p(\boldsymbol{x}_k|\widehat{\boldsymbol{\psi}}_x), \tag{3.11}$$

with $\widehat{\boldsymbol{\psi}}_x$ taken (for example) to be the posterior median, then

$$\boldsymbol{z}^K|\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}} \sim N(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}), \tag{3.12}$$

where the estimated mean and covariance are functions of $\widehat{\boldsymbol{\psi}}_x$. We could also incorporate the uncertainty in the posterior, use

$$\widehat{\boldsymbol{\mu}} = \mathrm{E}[\boldsymbol{z}_{kj}|\boldsymbol{x}^M], \quad \widehat{\boldsymbol{\Sigma}} = \mathrm{cov}(\boldsymbol{z}_{kj}|\boldsymbol{x}^M),$$

and replace the normal form in (3.12) with a Student's t distribution, bringing this stage of the model close to that of Zidek *et al.* (1998). Although the distribution given by (3.12) is high dimensional, the moments can be determined in an initial analysis, greatly reducing the computational burden. Note that (3.12) represents a Berkson error model with heteroscedastic errors.

Stage 3, Prior distributions. Specify priors for $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$.

4. SIMULATION STUDY

We now describe a simulation study with two aims. The first is to investigate the use of the convolution model (3.3), and the second is to assess the impact of estimated, rather than known, exposures in the health model. The location and observed exposures from the air pollution monitors, and the ED locations and populations at risk, are based on the London study described in Section 1.

The overall structure of the simulation is as follows: We fit a GRF model to the 16 observed monitor exposures and then, based on the estimated parameters, we simulate 1027 exposures at each of the study ED centroids, and also at the 16 monitor sites. In Section 4.1, the 1027 values are taken as the known exposures, and we compare the individual, convolution, and ecological models. In Section 4.2, we fit a GRF to the 16 simulated monitor values, and then obtain predictions at each of the 1027 ED centroids, in order to investigate the use of estimated exposures. We would expect the measured exposure to be most appropriate for those individuals living close to a monitor, and so we consider different designs in which the study population consist of individuals lying within $0.1r$ km of each of the 16 pollution monitors, for $r = 2, \ldots, 10$.

Letting x_m denote the measured SO_2 at monitor m, we take $\log x_m$, $m = 1, \ldots, 16$, as arising from an isotropic GRF model with mean μ_x, and covariance function of observations at locations $\boldsymbol{s}$ and $\boldsymbol{s}'$, $\sigma_x^2 \exp(-\phi_x|\boldsymbol{s} - \boldsymbol{s}'|)$, so that $\boldsymbol{\psi}_x = (\mu_x, \sigma_x^2, \phi_x)$. We used the 'GeoBUGS' software (Thomas *et al.*, 2000) to fit a GRF model with priors taken as: μ_x improper uniform, $\sigma_x^{-2} \sim \text{Ga}(0.01, 0.001)$, and $\phi_x \sim \text{U}(0.12, 1.15)$. The gamma prior is quite flat, while the prior for ϕ_x allows both very weak and very strong spatial dependence, relative to the study geography/monitor configuration (see Thomas *et al.*, 2000, for more details). We obtained posterior median estimates of $\widehat{\mu}_x = 0.69$, $\widehat{\sigma}_x = 0.71$, and $\widehat{\phi}_x = 0.84$.

4.1 *Known exposures*

We simulated disease counts for each ED, based on a log-linear Poisson model with a relative risk of 2. We report a single simulation only, but set $\beta_0 = -5.5$ which gives sufficient cases for a repeat simulation to give very similar results. Three models were fitted: the individual model (3.6), with 1027 pairs (Y_{kj}, x_{kj}), $j = 1, \ldots, m_k$; the ecological model (2.1), with 16 pairs (Y_k, x_k); and the convolution model (3.3), with 16 counts Y_k, and the 1027 exposures x_{kj}, with N_{kj} being taken as the true ED populations, $j = 1, \ldots, m_k$, $k = 1, \ldots, 16$.

Table 3 summarizes the results over different study radii. For radii greater than 200 m, we see unbiased estimation for the individual and convolution models, while the ecological model underestimates β_1. The individual and convolution model estimates are similar, with larger standard errors for the latter, reflecting the loss in information when there is no linkage between outcome and exposure (as we saw in Section 3.3). In the limit ($m_k = 1$, with a single exposure, for $K = 16$ areas) the three models would provide identical inference. The inaccuracy of estimation for a radius of 200 m is due to sampling variability (there are only 50 cases). Figure 2 shows individual, ecological, and convolution profile log-likelihoods for β_1. The loss in information in moving from the individual to the convolution designs is clear, as is the bias in the ecological estimator, which reduces as the study region diminishes in size (though sampling variability dominates at 200 m). The quadratic shape of the log-likelihoods indicates that asymptotic inference via quasi-likelihood is accurate.

For the individual model $\widehat{\kappa} \approx 1$, while for the convolution model it is slightly larger than unity. The estimated overdispersion is much larger for the ecological model, reflecting model misspecification: the ecological responses do not follow the Poisson model (2.1).

4.2 *Estimated exposures*

In this section, we repeat the fitting of the individual and convolution models, but now use estimated exposures. The simulated exposure data at the 16 monitors were analyzed with a GRF model, resulting in the estimates $\widehat{\mu}_x = 0.61, \widehat{\sigma}_x = 0.69$, and $\widehat{\phi}_x = 0.85$. We obtained predictions at each of the 1027 ED centroids to give our estimated exposures. The first approximation described in Section 3.4 was used for inference.

Table 4 gives the results; the ecological model results are identical to Section 4.1 but are included for completeness. Figure 3 shows individual, ecological, and convolution profile log-likelihoods for β_1. The horizontal axes are the same as in Figure 2, but the vertical axes differ.

Table 3. *Simulated data with known exposures: estimation for different study radii and different models; the true value of the log-relative risk, β_1, is log* $2 = 0.69$

Radii (km)	Number of EDs	Population size	Number of cases	Individual model			Ecological model			Convolution model		
				$\widehat{\beta}_1$	s.e.$(\widehat{\beta}_1)$	$\widehat{\kappa}$	$\widehat{\beta}_1$	s.e.$(\widehat{\beta}_1)$	$\widehat{\kappa}$	$\widehat{\beta}_1$	s.e.$(\widehat{\beta}_1)$	$\widehat{\kappa}$
1.0	1027	49 264.6	1674	0.67	0.010	1.0	0.46	0.098	55.8	0.68	0.014	1.3
0.9	847	40 141.2	1362	0.66	0.012	1.0	0.46	0.083	32.5	0.67	0.017	1.4
0.8	682	32 274.7	1071	0.67	0.016	1.0	0.44	0.070	18.1	0.68	0.020	1.1
0.7	519	24 524.8	816	0.67	0.019	1.0	0.48	0.060	10.5	0.68	0.022	0.98
0.6	381	18 145.9	619	0.66	0.022	1.1	0.52	0.053	5.9	0.67	0.028	1.3
0.5	265	12 503.4	396	0.68	0.027	1.1	0.55	0.054	3.9	0.69	0.024	0.63
0.4	169	8067.9	215	0.70	0.032	0.99	0.61	0.063	2.9	0.71	0.033	0.89
0.3	96	4863.0	114	0.64	0.047	1.0	0.63	0.063	1.4	0.64	0.053	1.2
0.2	40	2135.8	50	0.59	0.099	0.88	0.46	0.069	0.65	0.59	0.090	0.68

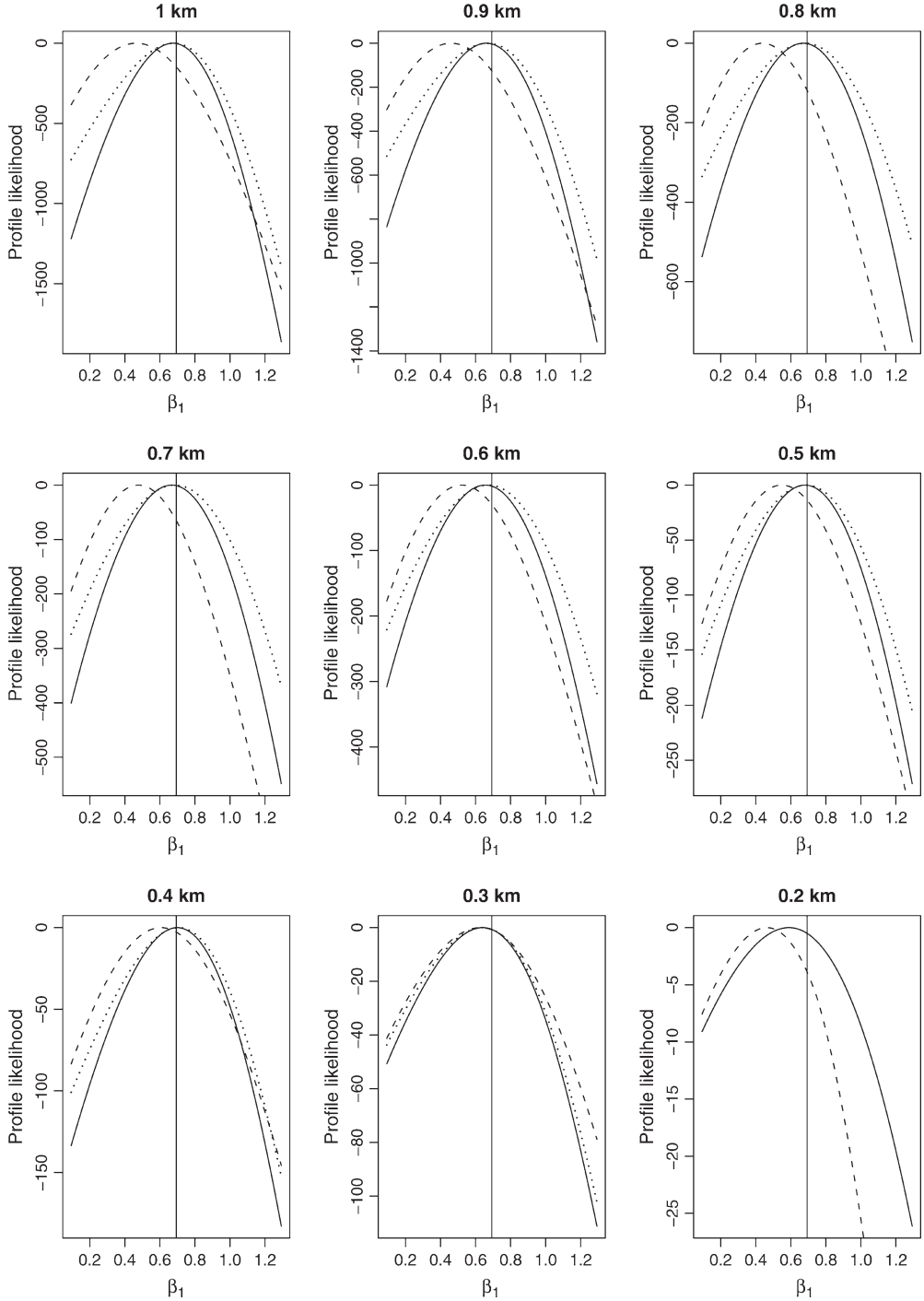

Fig. 2. Profile log-likelihoods for simulated data with known exposures. The solid, dashed, and dotted lines correspond to individual, ecological, and convolution models, respectively. The solid vertical line on each plot represents the true value of β_1.

Table 4. *Simulated data with estimated exposures: estimation for different study radii and different models; the true value of the log-relative risk, β_1, is log 2 = 0.69*

Radii (km)	Number of EDs	Population size	Number of cases	Individual model			Ecological model			Convolution model		
				$\widehat{\beta}_1$	s.e.($\widehat{\beta}_1$)	$\widehat{\kappa}$	$\widehat{\beta}_1$	s.e.($\widehat{\beta}_1$)	$\widehat{\kappa}$	$\widehat{\beta}_1$	s.e.($\widehat{\beta}_1$)	$\widehat{\kappa}$
1.0	1027	49 264.6	1674	1.04	0.069	6.5	0.46	0.098	55.8	1.19	0.16	28.6
0.9	847	40 141.2	1362	0.99	0.058	4.0	0.46	0.083	32.5	1.09	0.13	16.6
0.8	682	32 274.7	1071	0.93	0.051	2.6	0.44	0.070	18.1	0.99	0.10	8.8
0.7	519	24 524.8	816	0.91	0.051	2.2	0.48	0.061	10.5	0.97	0.087	5.7
0.6	381	18 145.9	619	0.90	0.057	2.2	0.52	0.053	5.9	0.96	0.077	3.7
0.5	265	12 503.4	396	0.88	0.068	2.3	0.55	0.054	3.9	0.92	0.083	3.2
0.4	169	8067.9	215	0.90	0.087	2.4	0.61	0.063	2.9	0.93	0.092	2.6
0.3	96	4863.0	114	0.85	0.086	1.4	0.63	0.063	1.4	0.85	0.089	1.5
0.2	40	2135.8	50	0.59	0.11	0.91	0.46	0.069	0.65	0.59	0.092	0.67

Both the individual and convolution models produce $\widehat{\beta}_1$ with positive bias because the 16 observed exposures are not sufficient to characterize the exposure surface. This is illustrated in Figure 4 in which the 'known' exposures are plotted versus distance from the closest monitor for two representative sites on the top row, with the 'estimated' exposures on the bottom row. The modeled exposures for London Bridge Place are not only determined by the concentration measured at that site (2.64 ppb) and the overall mean (2.86 ppb) but also increased due to the high exposure measured at Cromwell Road (4.38 ppb), which is only 2.5 km away (sites 5 and 6 on Figure 1). We see that the estimated exposures do not reflect the variability of the true exposures, and exhibit the well-known attenuation to the overall mean of these shrinkage-type estimators. This attenuation results in a narrowing of the estimated exposure range as compared to the true exposure range, resulting in overestimation of the regression coefficient.

For radii of 300–500 m, the ecological analysis actually provides more accurate estimation than the individual and convolution models, since it is based on known exposures, albeit at just 16 points. Hence, it may actually be detrimental to model the exposure surface. The very large values of $\widehat{\kappa}$ indicate the difficulties in estimation of the exposures (though in practice one would not know whether this overdispersion was due to other problems, such as missing confounders, and/or misrecording of population/health counts).

Analyses using the three-stage model of Section 3.4 are not reported. The results were poor because the simple errors-in-variables model (3.12) cannot correct for the attenuation problems discussed above. Future research will examine when the three-stage model provides accurate inference, in particular as a function of the spatial density of monitor information, relative to the exposure variability.

5. Mortality and SO_2 in London

We now return to the example introduced in Section 1. We carried out a number of simulations, similar to those of the previous section, but found that for the observed number of cases the results were highly variable; hence, we conclude that the observable exposure data are not sufficient to reliably estimate the association in this study. We carried out individual, ecological, and convolution analyses, as in Section 4.2, but do not include the results since they are dominated by sampling variability. If there were more cases, then we might hope to see some correspondence between ecological and convolution analyses, at least for small radii. There is no benefit in using a simple errors-in-variables approach to correct for the estimated exposures since the pollution monitors are too sparsely located (relative to the spatial variability in exposure) to give reliable estimation of the log SO_2 surface.

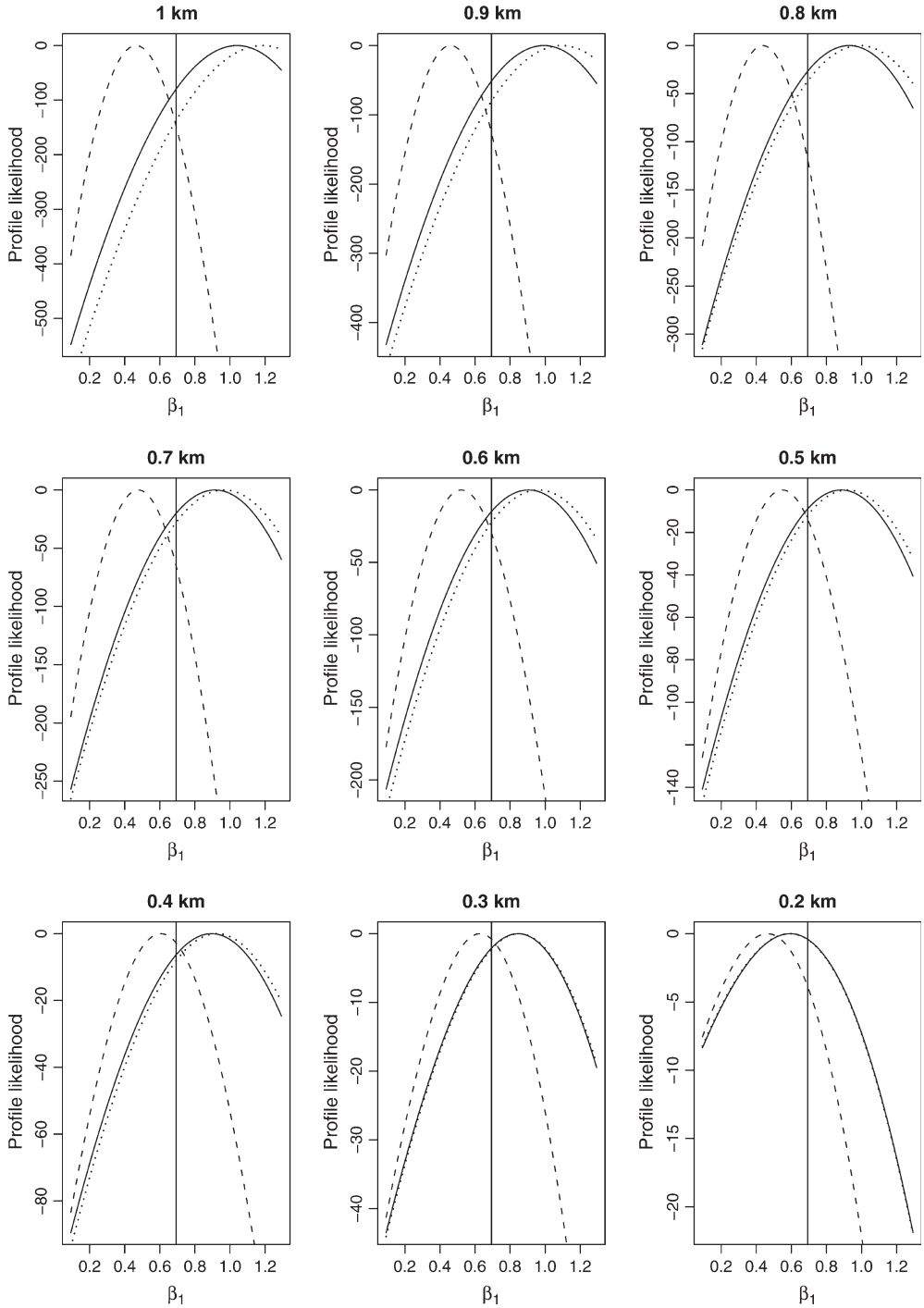

Fig. 3. Profile log-likelihoods for simulated data with estimated exposures. The solid, dashed, and dotted lines correspond to individual, ecological, and convolution models, respectively. The solid vertical line on each plot represents the true value of β_1.

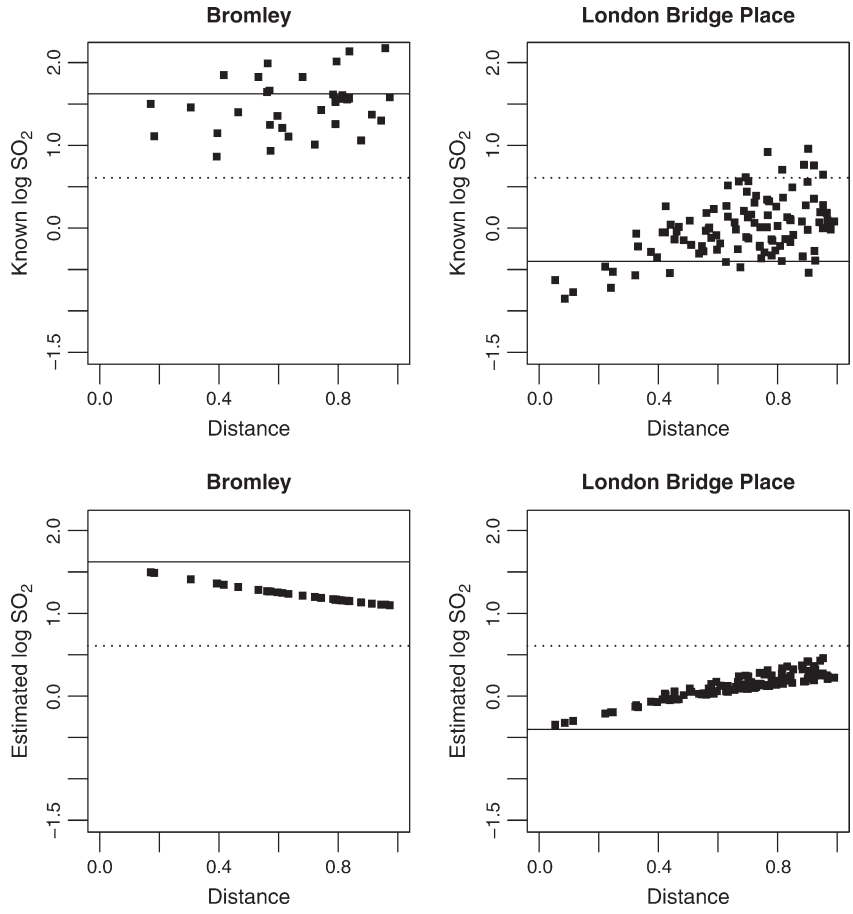

Fig. 4. Known (top row) and estimated (bottom row) log SO_2 versus distance from monitor for simulated data, for the Bromley and London Bridge Place monitors. On each plot, the dotted line represents the overall mean of the fitted GRF surface, and the solid line corresponds to the value of the log exposure at the monitor (located at at a distance of 0 km).

The analyses we carried out were based on assuming that all populations are located at their population ED centroids, and are therefore susceptible to ecological bias. Postcode centroids, which are far more dense geographically, were available, but we decided that refinement of the analysis to use this information was not merited, given that the exposure surface is so poorly estimated.

6. Discussion

In this paper, we have considered the common spatial epidemiological situation in which aggregate disease and population counts are available, along with exposure measures at a set of monitor sites. We have illustrated that a naive ecological regression is subject to pure specification bias, but have developed a convolution model that is not subject to bias, so long as accurate within-area exposure measures are available. A second conclusion is that estimated exposures should be used with caution since simple measurement error models cannot adjust for bias resulting from estimates based on sparse monitor

information. Some applications use a more complex exposure model, for example based on geographic and atmospheric variables (e.g. Briggs *et al.*, 1997), but the general point is that the predictions still need to be accurate.

The convolution model was derived with no confounding variables. We now describe a model in which we jointly estimate confounder and exposure effects. At the level of the individual, let Y_{kci} be the disease indicator of individual i in confounder stratum c and area k and assume

$$Y_{kci}|x_{kci}, \boldsymbol{\beta}, \boldsymbol{\gamma} \sim \text{Bernoulli}\{\exp(\beta_0 + \beta_1 x_{kci} + \gamma_c)\},$$

for $k = 1, \ldots, K$, $c = 1, \ldots, C$, $i = 1, \ldots, N_{kc}$. Usually the numbers of individuals and cases within each confounder stratum by area, N_{kc} and Y_{kc}, will be known. Aggregating over individuals within confounder stratum, and assuming a rare disease, gives

$$Y_{kc}|\boldsymbol{x}_{kc}, \boldsymbol{\beta}, \boldsymbol{\gamma} \sim_{\text{ind}} \text{Poisson}\left\{ e^{\beta_0+\gamma_c} \sum_{i=1}^{N_{kc}} \exp(\beta_1 x_{kci}) \right\},$$

where $\boldsymbol{x}_{kc} = \left(x_{kc1}, \ldots, x_{kcN_k}\right)$. Suppose now that we have exposure measurements x_{kcj} by confounder stratum, $j = 1, \ldots, m_k$, at m_k locations within area k. Then

$$Y_{kc}|\boldsymbol{x}_{kc}, \boldsymbol{\beta}, \boldsymbol{\gamma} \sim_{\text{ind}} \text{Poisson}\left\{ e^{\beta_0+\gamma_c} \sum_{j=1}^{m_k} N_{kcj} \exp(\beta_1 x_{kcj}) \right\},$$

where $\boldsymbol{x}_{kc} = \left(x_{kc1}, \ldots, x_{kcm_k}\right)$ and we take N_{kcj} as the confounder-defined populations in sub-area (e.g. ED) j. If we assume that individuals in the same sub-area are subject to the same exposure $\boldsymbol{x}_k = \left(\boldsymbol{x}_{k1}, \ldots, \boldsymbol{x}_{km_k}\right)$ regardless of confounder group, which means that we do not have a within-area confounder, then we obtain

$$Y_{kc}|\boldsymbol{x}_k, \boldsymbol{\beta}, \boldsymbol{\gamma} \sim_{\text{ind}} \text{Poisson}\left\{ e^{\beta_0+\gamma_c} \sum_{j=1}^{m_k} N_{kcj} \exp(\beta_1 x_{kj}) \right\}. \tag{6.13}$$

The assumption that the exposure distribution is the same across potential confounder stratum within areas is clearly crucial, and will need to be critically assessed in any application. For example, gender may be less related to exposure than age is since different age groups may have very different time activities and therefore exposure profiles.

We have concentrated upon inference via quasi-likelihood, but an obvious extension is to include random effects which allow for unmeasured variables with spatial structure. Clayton *et al.* (1993) used the model of Besag *et al.* (1991) in an ecological regression context, but did not allow for within-area variability in exposure. We extend (2.2) to the form

$$Y_{ki}|x_{ki}, \boldsymbol{\beta}, U_k, V_k \sim_{\text{ind}} \text{Bernoulli}\{\exp(\beta_0 + \beta_1 x_{ki} + U_k + V_k)\}, \tag{6.14}$$

where U_k and V_k represent random effects with and without spatial structure, retrospectively, in area k. Under aggregation model, (6.14) takes the form

$$Y_k|\boldsymbol{\beta}, \boldsymbol{x}_k, U_k, V_k \sim \text{Poisson}\left\{ e^{\beta_0+U_k+V_k} \sum_{j=1}^{m_k} N_{kj} \exp(\beta_1 x_{kj}) \right\}. \tag{6.15}$$

The inclusion of random effects cannot control for general confounding. Clayton *et al.* (1993) argue that spatial random effects can control for 'confounding by location',though this is difficult to achieve in

practice since regression estimates can be sensitive to the particular spatial model used. The model may provide more appropriate standard errors than under an assumption of independent outcomes, however. If there is evidence of residual spatial dependence, then we would recommend carrying out sensitivity analyses under a range of scenarios, including models that do and do not acknowledge spatial dependence.

Model (3.2), with some modification, also allows computation for the disease-mapping model of Kelsall and Wakefield (2002) to be carried out without recourse to approximation. For a related approach, see Follestad and Rue (2003).

A difficult yet crucial issue in any analysis that uses spatially referenced exposure data is whether to model the exposure surface. As an aid to making this decision, we would recommend following the procedure described in Section 4. Specifically, one may fit a model to the monitor exposure data, and simulate new monitor and population location exposure data using the fitted model; the differences between known and estimated values can then be examined, to gain insight into whether an exposure modeling strategy is likely to be successful. The study design will often inform the need to model the exposure surface.

An interesting design question is the determination of which populations to study in relation to the location of the pollution monitors. This choice represents the classic mean–variance trade-off; populations close to a monitor have accurate exposure assessment but are small in size, while examining larger populations gives an increase in power but results in less accurate exposure estimates. One way of increasing power is to have a dense monitoring network, where dense is relative to both the spatial variability in exposure and the population distribution. If the exposure surface is relatively flat, only a sparse network is required, but in this case a spatial study will have low power due to narrow exposure contrasts. In studies of the acute effects of air pollution, temporal contrasts provide the greatest exposure information, which suggests that in such a study, modeling spatial variability in exposure will not be worthwhile.

ACKNOWLEDGMENTS

The data analyzed in Section 5 were supplied by the Small Area Health Statistics Unit, a unit that is funded by a grant from the Department of Health; Department of the Environment, Food and Rural Affairs; Environment Agency; Health and Safety Executive; Scottish Executive; National Assembly for Wales; and Northern Ireland Assembly. This work of the first author was supported by grant R01 CA095994 from the National Institutes of Health. The focus of the article was greatly helped by the constructive comments of the editor and a referee. *Conflict of Interest:* None declared.

REFERENCES

BESAG, J., YORK, J. AND MOLLIÉ, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics* **43**, 1–59.

BRIGGS, D. J., COLLINS, S., ELLIOTT, P., FISCHER, P., KINGHAM, S., LEBRET, E., PRYL, K., VAN REEUWIJK, H., SMALLBONE, K. AND VAN DER VEEN, A. (1997). Mapping urban air pollution using GIS: a regression-based approach. *International Journal of Geographical Information Systems* **11**, 699–718.

CARLIN, B. P., XIA, H., DEVINE, O., TOLBERT, P. AND MULHOLLAND, J. (1999). Spatio-temporal hierarchical models for analyzing Atlanta pediatric asthma ER visit rates. In Gatsonis, C., Kass, R. E., Carlin, B., Carriquiry, A., Gelman, A., Verdinelli, I. and West, M. (eds), *Case Studies in Bayesian Statistics*, volume IV. New York: Springer, pp. 303–320.

CLAYTON, D., BERNARDINELLI, L. AND MONTOMOLI, C. (1993). Spatial correlation in ecological analysis. *International Journal of Epidemiology* **22**, 1193–1202.

DIGGLE, P. AND ELLIOTT, P. (1995). Disease risk near point sources: statistical analyses for analyses using individually or spatially aggregated data. *Journal of Epidemiology and Community Health* **49**, S20–S27.

ELLIOTT, P., ARNOLD, R., COCKINGS, S., EATON, N., JARUP, L., JONES, J., QUINN, M., ROSATA, M., THORNTON, I., TOLEDANO, M. *et al.* (2000). Risk of mortality, cancer incidence and stroke in a population potentially exposed to cadmium. *Occupational and Environmental Medicine* **57**, 94–97.

FOLLESTAD, T. AND RUE, H. (2003). Modelling spatial variation in disease risk using Gaussian Markov random field proxies for Gaussian random fields (submitted).

GELFAND, A. E., ZHU, L. AND CARLIN, B. P. (2001). On the change of support problem for spatio-temporal data. *Biostatistics* **2**, 31–45.

GREENLAND, S. (1992). Divergent biases in ecologic and individual studies. *Statistics in Medicine* **11**, 1209–1223.

GREENLAND, S. AND MORGENSTERN, H. (1989). Ecological bias, confounding and effect modification. *International Journal of Epidemiology* **18**, 269–274.

GREENLAND, S. AND ROBINS, J. (1994). Ecological studies: biases, misconceptions and counterexamples. *American Journal of Epidemiology* **139**, 747–760.

GUTHRIE, K. A. AND SHEPPARD, L. (2001). Overcoming biases and misconceptions in ecological studies. *Journal of the Royal Statistical Society, Series A* **164**, 141–154.

KELSALL, J. E. AND WAKEFIELD, J. C. (2002). Modeling spatial variation in disease risk: a geostatistical approach. *Journal of the American Statistical Association* **97**, 692–701.

LE, N. D., SUN, W. AND ZIDEK, J. V. (1996). Bayesian multivariate spatial interpolation with data missing-by-design. *Journal of the Royal Statistical Society, Series B* **59**, 501–510.

MAHESWARAN, R., MORRIS, S., FALCONER, S., GROSSINHO, A., PERRY, I., WAKEFIELD, J. AND ELLIOTT, P. (1999). Magnesium in drinking water supplies and mortality from acute myocardial infarction in north west England. *Heart* **82**, 455–460.

MCCULLAGH, P. AND NELDER, J. A. (1989). *Generalized Linear Models*, 2nd edition. London: Chapman and Hall.

OTT, W. R. (1994). *Environmental Statistics and Data Analysis*. Boca Raton, FL: Lewis Publishers.

PIANTADOSI, S., BYAR, D. P. AND GREEN, S. B. (1988). The ecological fallacy. *American Journal of Epidemiology* **127**, 893–904.

PLUMMER, M. AND CLAYTON, D. (1996). Estimation of population exposure. *Journal of the Royal Statistical Society, Series B* **58**, 113–126.

POPE, C. A. AND DOCKERY, D. (1996). Epidemiology of chronic health effects: cross-sectional studies. In Wilson, R. and Spengler, J. (eds), *Particles in Our Air: Concentrations and Health Effects*. Boston: Harvard University Press, pp. 149–167.

PRENTICE, R. L. AND SHEPPARD, L. (1995). Aggregate data studies of disease risk factors. *Biometrika* **82**, 113–125.

RICHARDSON, S., STUCKER, I. AND HÉMON, D. (1987). Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Epidemiology* **16**, 111–120.

SELVIN, H. C. (1958). Durkheim's 'suicide' and problems of empirical research. *American Journal of Sociology* **63**, 607–619.

SHADDICK, G. AND WAKEFIELD, J. C. (2002). Modelling multivariate pollutant data at multiple sites. *Applied Statistics* **51**, 351–372.

SHEPPARD, L. AND PRENTICE, R. L. (1995). On the reliability and precision of within- and -population estimates of relative rate parameters. *Biometrics* **51**, 853–863.

SIMPSON, S., DIAMOND, I., TONKIN, P. AND TYE, R. (1996). Updating small area population estimates in England and Wales. *Journal of the Royal Statistical Society, Series A* **159**, 235–247.

THOMAS, A., BEST, N. G., ARNOLD, R. A. AND SPIEGELHALTER, D. J. (2000). *GeoBUGS User Manual*. London: Imperial College of Science, Technology and Medicine.

TONELLATO, S. F. (2001). A multivariate time series model for the analysis and prediction of carbon monoxide atmospheric concentrations. *Applied Statistics* **50**, 187–200.

WAKEFIELD, J. C. (2003). Sensitivity analyses for ecological regression. *Biometrics* **59**, 9–17.

WAKEFIELD, J. C. (2004). Ecological inference for 2 × 2 tables (with discussion). *Journal of the Royal Statistical Society, Series A* **167**, 385–445.

WAKEFIELD, J. C. AND SALWAY, R. E. (2001). A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society, Series A* **164**, 119–137.

ZHU, L., CARLIN, B. P. AND GELFAND, A. E. (2003). Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER visits in Atlanta. *Environmetrics* **14**, 537–557.

ZIDEK, J. V., WHITE, R., LEE, N. D., SUN, W. AND BURNETT, R. T. (1998). Imputing unmeasured explanatory variables in environmental epidemiology with application to health impact analysis of air pollution. *Environmental and Ecological Statistics* **5**, 99–115.

[*Received July 2, 2005; first revision November 4, 2005; second revision December 21, 2005; accepted for publication January 10, 2006*]