Spatial Poisson Regression for Health and Exposure Data Measured at Disparate Resolutions

Nicola G. BEST, Katja ICKSTADT and Robert L. WOLPERT[1]

JASA Manuscript A98-156-R0
Revised March 27, 2000, 10:17

Ecological regression studies are widely used to examine relationships between disease rates for small geographical areas and exposure to environmental risk factors. The raw data for such studies, including disease cases, environmental pollution concentrations and the reference population at risk, are typically measured at a variety of levels of spatial aggregation but are accumulated to a common geographical scale to facilitate statistical analysis. In this traditional approach heterogeneous exposure distributions within the aggregate areas may lead to biased inference, while individual attributes such as age, gender and smoking habits must either be summarized to provide area level covariate values or used to stratify the analysis.

This paper presents a spatial regression analysis of the effect of traffic pollution on respiratory disorders in children. The analysis features data measured at disparate, non-nested scales, including spatially varying covariates, latent spatially varying risk factors, and case-specific individual attributes.

The problem of disparate discretizations is overcome by relating all spatially varying quantities to a continuous underlying random field model. Case-specific individual attributes are accommodated by treating cases as a marked point process. Inference in these hierarchical Poisson/gamma models is based on simulated samples drawn from Bayesian posterior distributions, using Markov chain Monte Carlo methods with data augmentation.

[1]Nicola G. Best is Lecturer in Biostatistics, Department of Epidemiology and Public Health, Imperial College School of Medicine, St Marys' Campus, London W2 1PG, UK (email: n.best@ic.ac.uk); Katja Ickstadt is Postdoctoral Research Fellow, Department of Mathematics, Darmstadt University of Technology, 64289 Darmstadt, Germany (email: ickstadt@mathematik.tu-darmstadt.de); and Robert L. Wolpert is Professor, Institute of Statistics and Decision Sciences, Duke University, Box 90251, Durham, NC 27708-0251 (email: wolpert@stat.duke.edu). The authors acknowledge gratefully the support of the Deutsche Forschungsgemeinschaft, of US NSF grants DMS-9626829 and DMS-9707914, of a Wellcome Research Travel grant, and of UK Medical Research Council grant G9803841. The SAVIAH study was funded by a grant from the European Union Third Framework Environment Programme. The authors thank Paul Elliott, Jeremy Bullard, David Briggs, Sue Collins, and Jon Wakefield for providing data and sharing insight, and Dave Higdon and the editor, associate editor and three anonymous referees for helpful suggestions.

KEY WORDS: Air pollution, Bayesian hierarchical models, Continuous random fields, Ecological regression, Environmental epidemiology, Respiratory illness.

1. INTRODUCTION

The potential health hazards posed by many environmental pollutants have led to concerns about the impact of traffic-related air pollution on respiratory illnesses in children (Wardlaw 1992). Childhood respiratory symptoms may also be associated with socio-economic deprivation and individual attributes such as gender, passive smoking due to parental tobacco use and aspects of the home environment like damp walls and the use of gas for cooking and heating (Elliott, Briggs, Lebret, Goryinski and Bohumil 1997; Fischer et al. 1998). Some of these may be thought of as attributable risk factors, while others modify individuals' susceptibility. Some are measured and reported in epidemiological studies, while other important factors may remain unobserved or unmeasurable. Many of these covariates vary smoothly geographically, leading to the expectation that disease incidence will also exhibit smooth spatial trends. The principal goal of geographical epidemiology is to increase our understanding of disease ætiology by revealing the relationship between disease incidence and these individual and spatially varying factors.

In this paper we explore the relationship between the prevalence of the respiratory ailment *severe wheeze* and road traffic pollution, adjusting for several individual and spatially varying risk factors. The data were collected as part of the European Small-Area Variations In Air quality and Health (SAVIAH) study. They include the residential postcode for reported severe wheeze cases in children aged seven to nine years living in Huddersfield, UK, along with individual attributes (age, gender, etc.) and home environment covariates (home dampness, maternal smoking, etc.) for each case. Postcodes in the UK cover a median of 12 households. The areas covered by a single postcode have no formal boundaries and can vary widely in size: in urban areas they are typically a few hundred meters across, although rural postcodes often cover a much larger area. Each postcode is georeferenced by a nominal centroid based on the easting and northing coordinates of the "first" residence in the postcode.

For traffic pollution levels we use annual average nitrogen dioxide (NO_2) concentrations available to us on a $250\,\mathrm{m} \times 250\,\mathrm{m}$ grid. Population density data, needed for studying population-based disease rates, were available from the 1991 UK census at the enumeration district (ED) level. The latter are administrative boundaries containing around 400 people on average. The geographical extent of EDs therefore varies considerably, ranging from a few hundred square meters in urban areas to several square kilometers in rural

EDs. Thus some of our data and covariates are given at the individual level and others at one of three non-nested group levels (postcode, grid or ED; see Figure 1).

Many previous studies have focused on the *acute* effects of exposure to air pollution in individuals (e.g., Schwartz 1994; Committee of Medical Effects of Air Pollution 1995a,b; Pope et al. 1995). These findings provide relatively clear evidence of a short-term association between respiratory illness and a range of pollutants. In contrast, the present data on NO_2 concentrations represent markers of *chronic* exposure to ambient air pollution. The potential adverse health effects of such exposures are vitally important because of the large numbers of people at risk. However, it is unclear to what extent, if any, the findings of acute studies can be extrapolated to predict the effects of chronic exposure to low levels of traffic-related outdoor air pollutants. A number of recent studies provide contradictory findings. For example, while several have reported positive associations between, variously, respiratory symptoms, hospitalization or lung function of children living close to roads compared to those living further away, or living in streets with high levels of traffic, self-reported truck-traffic or high predicted NO_2 levels (Murakami, Ono and Tamura 1990; Wjst et al. 1993; Edwards, Walters and Griffiths 1994; Weiland, Mundt, Ruckmann and Keil 1994; Pershagen, Rylander, Norberg, Eriksson and Nordell 1995; Duhme et al. 1996; Osterlee, Drijver, Lebret and Brunekreef 1996; Brunekreef, Janssen, de Hartog, Harssema and Knape 1997), others have reported no association (Waldron, Pottle and Dod 1995; Livingstone, Shaddick, Grundy and Elliott 1996; Briggs et al. 1997; Wilkinson et al. 2000).

Part of the explanation for the equivocal literature may relate to the methodological difficulties associated with many of the studies, including major problems of confounding and bias. There are two traditional approaches to exploring exposure/response relationships in epidemiology. One is *individual level* studies of case control or cohort design, typically using logistic or Poisson log-linear regression to relate disease incidence to individual level covariates and geographically-varying suspected environmental risk factors. Unfortunately this offers no way to discover and exploit geographical correlation and trends that might be associated with unreported spatially varying risk factors.

The second is *group level* or *small area* studies, in which *ecological regression* methods are used to relate apparent geographical variations in disease occurrence to the geographical distribution of suspected risk factors. Disease counts are aggregated over geographical units and are compared to aggre-

gated covariate summaries, often using lattice-based Gaussian or log-linear Poisson Markov random field (MRF) models to accommodate local spatial correlation (Clayton and Kaldor 1987; Cressie and Chan 1989; Besag, York, and Mollié 1991; Bernardinelli, Pascutto, Best, and Gilks 1997; Waller, Carlin, Xia, and Gelfand 1997). The group level approach offers the advantages of requiring only case counts and exposure averages over geographical areas (which are often available even when individual case and exposure data are not) and of revealing and exploiting spatial correlation in the data. However it has several serious shortcomings. One is that inference will depend on the geographical areas selected for analysis, for which there will be competing choices: larger units will exhibit less sampling variation but smaller ones will offer a more detailed view of geographical patterns in disease incidence; political or administrative areas are convenient for data analysis and presentation but vary widely in size and shape. This is often referred to as the *modifiable areal unit problem* in the geography literature (Openshaw 1984; Fotheringham and Wong 1991). Another is that the process of data aggregation will introduce approximation errors and uncertainty whenever (as in our case) covariate data are available at disparate non-nested levels of aggregation. A third is that the aggregation may distort or mask the true exposure/response relationship for individuals, a phenomenon called the *ecological fallacy*; this may occur for many reasons, for example, when exposure is inhomogeneous within areas and effects are nonlinear, or when a confounding factor leads to nonuniform baseline rates or modifies the exposure/response relationship across areas (Robinson 1950; Greenland 1992; Richardson 1992; Morgenstern 1998).

Evidently these two approaches have complimentary advantages: individual level studies permit individual level covariates to be treated as case-specific, without aggregating over groups in ways that affect and may distort inference, while small area studies are able to reflect geographical correlation and trends.

Our Bayesian spatial modeling approach combines the advantages of both by using each variable at its given level of spatial resolution and relating each quantity to an underlying continuous-parameter spatial random field. It reflects individual attributes at the case level, to reduce or eliminate the ecological fallacy; it captures spatial autocorrelation; and it scales appropriately to offer coherent inference at arbitrary levels of spatial resolution. The models are based on the Bayesian hierarchical point process models for dependent count data introduced by Wolpert and Ickstadt (1998) and extended in Ickstadt and Wolpert (1999) to accommodate individual level attributes

as marks; here we further extend the models to allow multiple non-nested levels of data aggregation and spatially varying explanatory variables. The SAVIAH data is introduced in Section 2., our inferential aim in the context of previous modeling approaches in Section 3., our model in Sections 4. and 5., and the results in Section 6. A discussion of the modeling approach and suggestions for future work appear in Section 7.

2. SMALL AREA VARIATIONS IN AIR QUALITY AND HEALTH: THE SAVIAH STUDY

Concerns about the health effects of outdoor air pollution from industrial sources and road traffic are currently the focus of considerable public and scientific attention. Indeed, reducing the harmful health impact associated with outdoor air quality is one of the issues raised in the UK government's *National Environmental Health Action Plan* (H.M. Government, 1996), drawn up under the auspices of the World Health Organization (WHO). However, studies of this relationship have been beset by methodological difficulties, including major problems of exposure estimation. The SAVIAH study (Elliott et al. 1997) was funded by the EU in 1993–94 to address some of the technical, methodological and substantive issues encountered in assessing the relation of air quality to health. The study was carried out in four field centers in the UK, Netherlands, Poland and the Czech Republic. Although similar methods were used in the four centers, our concerns about comparability of census data across national boundaries and results for questionnaires translated into different languages led us to base the present analysis solely on data from the UK center in Huddersfield.

2.1 Health Event and Population Data

Health outcomes were assessed using questionnaire surveys which were distributed to the parents or guardians of all children aged seven to nine years who attended a primary school in the study region. The questionnaire contained items on children's chronic respiratory symptoms, based on the WHO (Florey and Leeder 1973) and ISAAC (Pearce et al. 1993) standard parent-administered questionnaire. These are widely used to assess prevalence of childhood respiratory illness. The core questions relating to diagnosis of asthma are reported to have sensitivity of 0.68 and specificity of 0.82 to predict bronchial hyper-responsiveness, the latter being an objective marker of asthma (Shaw et al. 1992). Questions relating to potential risk factors such

as parental smoking, damp housing and socioeconomic factors etc. were also asked. Valid questionnaires were completed for 4437 children in Huddersfield, giving a 90.1% response rate. The outcome of interest for the present study was the respiratory illness *severe wheeze*. Children were deemed to have a positive response if they had suffered from more than twelve attacks of wheeze in the last year *or* had at least one attack which was severe enough to limit speech (Strachan and Carey 1995). Figure 1 (a) shows the location of the centroid of the home postcode for all 191 cases identified in the study region. Five postcodes contain two cases each and the remaining 181 cases each have a unique postcode; the study region contains a further 5967 postcodes with no cases. Under the null hypothesis of uniform risk, the expected distribution of cases would be proportional to the population of seven to nine year old children at risk. This is indicated in Figure 1 (a) by polygons representing census EDs shaded according to the logarithmic population density of children aged 0–14 years in the 1991 UK census, scaled by a factor of 3/15 to adjust for the difference of age ranges. We use pro-rated census figures as our reference population, rather than simply the cohort counts, because the uneven distribution of the SAVIAH cohort leaves large areas with no recorded cohort members and many others with only one or two children. This would make it difficult or impossible to produce sensible risk estimates for these areas and would introduce artifactual geographical patterns that reflect variations in demographics rather than risk. Instead, our pro-rated census figures may be viewed as an estimate of the typical or average population of seven to nine year old children living in each ED.

2.2 Environmental Exposure and Other Covariate Data

Nitrogen dioxide concentration was chosen as a summary measure for traffic-related air pollution since it arises primarily from vehicle emissions and can be measured reliably with passive sampling. Pollution concentrations were initially monitored using a dense network of 80 fixed and 40 portable monitoring sites located throughout Huddersfield, each with twin Palmes diffusion tubes as passive samplers, during four two-week surveys conducted in 1993–94. Briggs et al. (1997) compared a range of techniques (trend surface analysis, Kriging, spline interpolation, dispersion models) but found a multiple regression on such factors as nearby land use and traffic volume to offer the best predictive power for extrapolating the measured NO_2 concentrations to the entire study region. They report that the extrapolation explains 82% of the variability in observed NO_2 concentrations at ten addi-

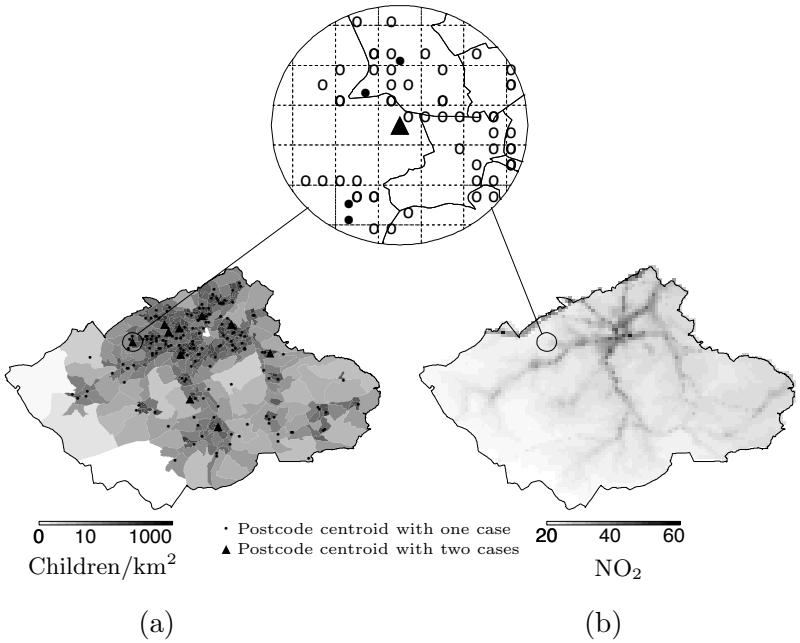


Figure 1. (a) Severe wheeze cases and population density in study region and (b) NO_2 concentrations in study region. 'Blow-up' circle shows enlargement of part of the study region to illustrate the three non-nested spatial scales: postcode centroid (including centroids of postcodes with no cases, denoted by unfilled circles), grid and ED.

tional out-of-sample monitoring sites reserved for model validation (further details may also be found in Best, Ickstadt, Wolpert, and Briggs (2000)). Concentrations were made available to us on the 250 m × 250 m grid shown in Figure 1 (b). Note that the dark streaks, indicating high NO_2 concentrations, tend to coincide with main roads and motorways.

Individual level data were collected in the questionnaire on each child's age, gender, home and school post codes, and features of the home environment including indicators of the use of gas for cooking or heating, signs of dampness or mould, presence of pets with fur or feathers, maternal smoking, or noise from nearby traffic. Any family history of respiratory ailments was also noted. A small number of these covariate data were missing due to non-response.

3. PREVIOUS APPROACHES TO MODELING THE SAVIAH DATA

Subsets of the SAVIAH data have been studied using a variety of models. Martuzzi and Elliott (1996) aggregate counts of cases of *mild* wheeze within school catchement areas (districts) and use empirical Bayes methods to produce smoothed estimates of disease rates, without including covariate information. They found significant heterogeneity in prevalence across the 71 school catchement areas, but no evidence to suggest that this variation was spatially correlated. Elliott et al. (1997) exploit the case-control design of the SAVIAH study to carry out individual level logistic regression analyses of mild wheeze and cough. After adjusting for gender, passive smoking, dampness in the home and other covariates, they found no evidence of an association with NO_2 concentration at the postcode of residence, although children whose parents reported being disturbed by traffic noise and fumes in the home were found to be at significantly higher risk of cough. In a preliminary analysis we adopted the hierarchical Bayesian approach of Besag et al. (1991) to carry out an area level generalized linear regression analysis of severe wheeze and traffic-related air pollution. The numbers of cases in EDs were treated as Poisson random variables whose log means depend linearly on both the average NO_2 concentration in the ED and a spatially correlated Gaussian MRF area level random effect intended to account for possible confounding due to unmeasured spatially correlated risk factors. This approach revealed a weak positive association (but with wide 95% credible limits including a relative risk of 1) between severe wheeze and average NO_2 concentration at the ED level.

Each of these approaches has inherent limitations. The absence of covariates in the model by Martuzzi and Elliott (1996) makes inference awkward. The use of a single point estimate of home NO_2 concentration by Elliott et al. (1997) cannot capture fully each child's true exposure to traffic pollution (no allowance was made for cumulative exposure over the local neighborhood of both the child's home and school, for example), and no provision is made for spatially correlated but unmeasured risk factors. The aggregation of NO_2 concentrations to the ED level in our preliminary area level analysis may distort individual exposure and may lead to ecological bias.

The approach we present below overcomes these limitations by using the case locations and all covariates at their given level of spatial aggregation and by treating the unknown true disease intensity and the spatially varying covariates as continuous random fields.

4. AN IDENTITY-LINK POISSON REGRESSION ANALYSIS

We view the 191 locations $\{y_n\}$ of severe wheeze cases in the SAVIAH data set and the vectors $\{a_n\}$ of individual attributes (age, gender, etc.) as the coordinates of "marked" points $\{x_n\}_{1\leq n\leq 191}$ in the product space $\mathcal{X} \equiv \mathcal{Y} \times \mathcal{A}$ of ordered pairs $x = (y, a)$ of locations $y \in \mathcal{Y}$ and attribute vectors (marks) $a \in \mathcal{A}$. We model the $\{x_n\}$ as the events of a Poisson random field $N(dx) \sim \mathsf{Po}\big(\Lambda(dx)\big)$ on $\mathcal{X}$ with uncertain and nonuniform intensity measure $\Lambda(dx)$, making $N(dx)$ a doubly-stochastic Poisson (Cox) process. The intensity measure $\Lambda(dx)$ reflects both known spatially varying features (population density, environmental risk factors, etc.) and individual level risk factors (gender, passive smoking, home dampness, etc.) that may affect disease rates. In this section we describe in more detail the formulation of $\Lambda(dx)$ and our approach to inference about its relation to individual attributes and spatially varying covariates; in Section 5. below we will extend this to include any unknown (latent) spatially varying factors.

We take $\Lambda(dx) = \Lambda(x)\, w(dx)$ to have a density function $\Lambda(x)$ with respect to a reference weight measure $w(dx)$ intended to reflect the known distribution of the population and their attributes. Without loss of generality we can write $w(dx) = w_{\mathrm{Y}}(dy)w_{\mathrm{A}}(da|y)$ as the product of a nonnegative spatial weight measure $w_{\mathrm{Y}}(dy)$ (we use the population per ED in Huddersfield of children aged seven to nine years, in hundreds of children) and a location-specific conditional probability distribution measure $w_{\mathrm{A}}(da|y)$ of attributes. Here we take $w_{\mathrm{A}}(da|y)$ to be the empirical attribute distribution $w_{\mathrm{A}}(da)$

based on the entire SAVIAH cohort and treat it as location-independent (i.e., $w_{\mathrm{A}}(da|y_1) = w_{\mathrm{A}}(da|y_2)$ for $y_1 \neq y_2$). This gives $\Lambda(x) = \Lambda(y, a)$ the interpretation of a conditional disease rate, measured in cases per hundred population with attribute value a at location y. The overall disease rate in cases per hundred at location y is given by the integral $\int_{\mathcal{A}} \Lambda(y, a)\, w_{\mathrm{A}}(da|y)$.

4.1 Excess and Relative Risk Factors

The causal mechanisms underlying many diseases remain uncertain and different covariates may be expected to affect the disease incidence $\Lambda(x)$ in different ways. One general scenario is to assume a 'baseline' number of cases which would occur anyway; a number of extra cases may arise which are directly attributable to exposure to certain covariates, while other covariates may affect individuals' susceptibility to other agents. Breslow and Day (1980, chap. 2) refer to these as *excess* and *relative* risk factors, respectively, and discuss the use of additive and multiplicative models to describe their respective influence. Most epidemiological studies treat all covariates as multiplicative relative risks, in part because their log-linear models offer no alternative. This would be appropriate if each risk factor were viewed as making a subject more susceptible to each other factor, but not if the risk factors are viewed as contributing incrementally to the risk. The present models offer investigators a choice (albeit a subjective one) by allowing covariates to enter both additively and multiplicatively in the same model. In our application we treat home dampness, passive smoking, and environmental air pollution as additive excess risks, and treat gender as a multiplicative relative risk (being male does not cause severe wheeze, even though boys appear to be more susceptible than girls to the influence of other risk factors which do).

Excess risk factors enter the model for disease intensity additively, giving an identity link Poisson regression model of the form $\Lambda(x) = \sum X_j(x)\beta_j$ for risk factors $X_j(x)$ which may depend on either or both the location y or individual attribute vector a in $x = (y, a)$, and uncertain regression coefficients β_j. Multiplicative relative risk factors affect the scale of $\Lambda(x)$, suggesting a model of the form $\Lambda(x) = \Lambda_0(x) \exp\left(\sum X_j(x)\beta_j\right)$. In our identity link setting, additive excess risk factors X_j and their coefficients β_j must be nonnegative, and so cannot be protective, but multiplicative relative risks may enter on the logarithmic scale with either sign. We include both in the form

$$\Lambda(x) = c(\beta, y)\, e^{\sum_{j \in J_{\mathrm{M}}} X_j(x)\beta_j} \sum_{j \in J_{\mathrm{A}}} X_j(x)\beta_j \tag{1}$$

where J_{A} and J_{M} index the additive excess and multiplicative relative risks, respectively, and $c(\beta, y) \equiv \left\{ \int_{\mathcal{A}} \exp\left[\sum_{j \in J_{\mathrm{M}}} X_j(x)\beta_j \right] w_{\mathrm{A}}(da|y) \right\}^{-1}$ is a normalizing term leading to a mean relative risk factor of unity.

For our Bayesian analysis we model uncertainty about the coefficients $\{\beta_j\}$ through a joint prior distribution, described below.

4.2 Regression Results

Let $J_{\mathrm{M}} = \{0\}$ index our single multiplicative relative attribute, gender, with $X_0 = 1$ for boys and $X_0 = 0$ for girls, and let $J_{\mathrm{A}} = \{1, 2, 3, 4\}$ index a baseline risk or intercept term $X_1 \equiv 1$ and the three additive excess risks: binary indicators X_2 and X_3 of dampness and passive smoking in the home, respectively, and excess NO_2 concentration X_4 measured in $\mu g/m^3$ above the nominal baseline of 21 $\mu g/m^3$ (the minimum NO_2 concentration observed in the data).

For reference measure $w_{\mathrm{Y}}(dy)$ we use the estimated population (in hundreds) of seven to nine year old children, based on pro-rated 1991 UK census figures for children aged 0–14 (we do not model measurement error, population changes, etc). We take $w_{\mathrm{A}}(da|y)$ to be location-independent and use the mean of the empirical attribute distribution based on the SAVIAH cohort. Under this model the case density at site $y \in \mathcal{Y}$ with attribute vector $a \in \mathcal{A}$ is given by

$$\Lambda(x) = \frac{e^{X_0(x)\beta_0}}{(1 - \bar{X}_0) + \bar{X}_0\, e^{\beta_0}} \sum_{j=1}^{4} X_j(x)\beta_j \tag{2}$$

where $x = (y, a)$ and $\bar{X}_0 \equiv \int_{\mathcal{A}} X_0(x)\, w(dx)/w(\mathcal{X}) = .488$ is the SAVIAH cohort average of the Bernoulli variable $X_0(x)$, i.e., the population proportion of boys, so the normalizing term from Equation 1 is $c(\beta, y) \equiv \left\{ \int_{\mathcal{A}} \exp[X_0(x)\beta_0]\, w(dx) \right\}^{-1} = \{(1 - \bar{X}_0) + \bar{X}_0 e^{\beta_0}\}^{-1}$. For example, consider the two cases of severe wheeze marked by a triangle in the inset shown in Figure 1. Child 1 is a boy: there is no dampness in his house and his mother smokes (i.e., his individual attribute vector a takes values $X_0 = 1, X_2 = 0, X_3 = 1$); child 2 is a girl: there is no dampness in her house and her mother is a non-smoker (i.e., her individual attribute vector a takes

values $X_0 = 0, X_2 = 0, X_3 = 0$). The baseline risk term is included for both children ($X_1 = 1$). Denoting the centroid of their home postcode by y_*, and the value of the excess NO_2 concentration estimated at this location by $NO_2(y_*)$, the probability of child 1 having severe wheeze is given by

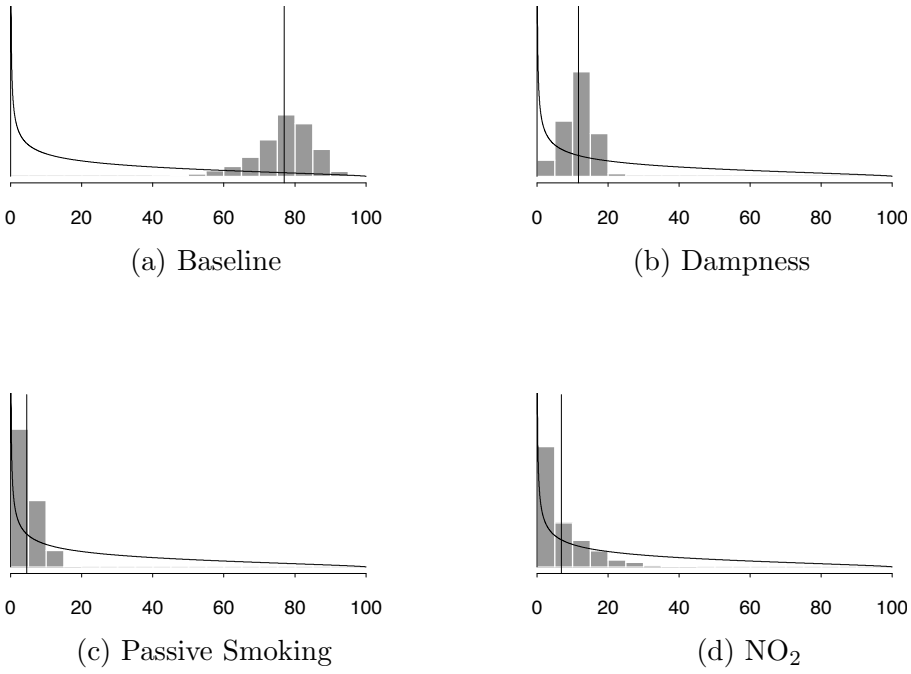
$$\Lambda(y^*, a = \{1, 0, 1\}) \;=\; \frac{e^{\beta_0}}{(1 - 0.488) + 0.488\, e^{\beta_0}} \{\beta_1 + \beta_3 + \mathrm{NO}_2(y_*)\beta_4\}$$

and the probability of child 2 having severe wheeze is given by

$$\Lambda(y^*, a = \{0, 0, 0\}) \;=\; \frac{1}{(1 - 0.488) + 0.488\, e^{\beta_0}} \{\beta_1 + \mathrm{NO}_2(y_*)\beta_4\}\,.$$

We used independent prior distributions for the regression coefficients β_j expressing only vague prior information. A standard normal was chosen for β_0 expressing prior ignorance as to whether boys would be more or less susceptible than girls and approximately 90% prior probability that their susceptibility ratio would be between one-fifth and five. Gamma distributions $\mathsf{Ga}(\alpha_j, \tau_j)$ were used for the four (necessarily positive) additive excess risk coefficients $\beta_1, ..., \beta_4$, with means $\mu_j = \alpha_j/\tau_j$ chosen to assign approximately equal prior expected risk $\bar{X}_j\, \mu_j = 1$ (i.e., one case per hundred population) to each factor, and shape parameters $\alpha_j = .575$ chosen to give each approximately 90% prior probability of lying between one tenth and ten times the prior mean. Here $\bar{X}_j \equiv \int_{\mathcal{X}} X_j(x)\, w(dx)/w(\mathcal{X})$ is the average within the SAVIAH cohort; this leads to inverse scale parameters $\tau_j = \alpha_j\, \bar{X}_j$ with values of $\tau_1 = .575$ for the baseline term (since $\bar{X}_1$=1), $\tau_2 = .148$ for dampness ($\bar{X}_2$=0.257), $\tau_3 = .154$ for passive smoking ($\bar{X}_3$=0.268), and $\tau_4 = 5.098$ for NO_2 ($\bar{X}_4$=8.860).

Markov chain Monte Carlo (MCMC) methods were used to obtain a sample from the joint posterior distribution of the $\{\beta_j\}$ (see Appendix A). Missing individual attributes a_j were sampled from their empirical distributions. Figure 2 shows the marginal prior densities for the four additive excess risk parameters $\beta_1, \ldots, \beta_4$, scaled as percent of total risk ($100\% \times \bar{X}_j\beta_j / \sum_i \bar{X}_i\beta_i$), along with posterior histograms and posterior means for these risk percentages. It shows that the model attributes most of the risk (77.0%) to the baseline risk term, and attributes about 6.9% to the total effects of traffic pollution (or, from Table 2, about .01 cases per 100 population at risk per $\mu g/m^3$ excess NO_2). The coefficient β_0 of the indicator for boys in the log relative risk factor has a posterior mean of .61 and posterior variance of .02, indicating that boys experience a relative risk about 1.85 times higher than that for girls. More details are given in Tables 1 and 2.



(a) Baseline (b) Dampness

(c) Passive Smoking (d) NO_2

Figure 2. Prior and posterior distributions of risk attribution (in percent) for the identity link model without a latent spatial covariate.

5. A RANDOM FIELD APPROACH

The simple regression approach above attributes less than one quarter of the risk, and therefore of the cases, to the included covariates, leaving the baseline risk term to "explain" the majority of cases. This term reflects both the baseline rate of severe wheeze amongst seven to nine year old children and also any unattributed risk due to unmeasured or unknown covariates not included in the model. We now explore whether there are spatial patterns in the distribution of that unattributed risk, by introducing a new latent spatial covariate $X_5(x)$.

If the unattributed risk should arise from a number of point sources of differing magnitudes γ_i at locations s_i in some set $\mathcal{S}$, and fall off with increasing distance from those sources at rates described by a kernel $k(x, s)$, e.g. a Gaussian kernel, the aggregate effect at location y for an individual with attribute vector a would be the kernel mixture $\sum k(x, s_i)\,\gamma_i$, given in integral form by

$$X_5(x)\beta_5 = \sum k(x, s_i)\,\gamma_i = \int_{\mathcal{S}} k(x, s)\Gamma(ds). \tag{3}$$

Here, $\Gamma(ds) = \sum \gamma_i \delta_{s_i}(ds)$ is a random field, where δ_s denotes the unit point mass at location s and $x = (y, a)$ denotes a marked point in $\mathcal{X} = \mathcal{Y} \times \mathcal{A}$. The term $X_5(x)\beta_5$ may be interpreted as a smoothly varying spatial random effect representing the underlying unattributed risk.

Gamma random fields have just such a representation as a discrete sum of (infinitely many) point masses, offering a convenient way to model uncertainty about the locations and magnitudes of the risk sources. Following Wolpert and Ickstadt (1998) we use a Gamma random field $\Gamma(ds) \sim \mathsf{Ga}\big(\alpha(ds), \tau(s)\big)$ with shape measure $\alpha(ds)$ and inverse scale function $\tau(s)$ leading to a Dirichlet process mixture for the latent spatial covariate $X_5(x) \equiv \int_{\mathcal{S}} k(x, s)\Gamma(ds)/\Gamma(\mathcal{S})$ and a Gamma distribution for the uncertain coefficient $\beta_5 \equiv \Gamma(\mathcal{S})$ (see Appendix B for a more detailed discussion on the Gamma random field). For our prior we used a uniform shape measure $\alpha_5(ds)$ on a rectangular study region $\mathcal{S}$ surrounding Huddersfield and a constant inverse scale $\tau_5(s)$. The value of $\tau_5(s)$ was chosen so that β_5 will have a Gamma distribution with the same shape parameter $\alpha_5(\mathcal{S}) \approx .575$ and $X_5(x)\beta_5$ will have the same prior mean $\alpha_5(\mathcal{S})/\tau_5(s) \approx 1$ as for the non-latent covariates of Section 4.2. We use a Gaussian kernel $k(x, s) = (\theta^2/2\pi)\exp(-\theta^2\,|y - s|^2/2)$ (recall $x = (y, a)$) with uncertain precision parameter θ, which may be interpreted as the inverse distance (in meters) over which the latent point sources exert influence.

A lognormal prior with median 1/400m and an 80% prior probability interval between 1/1500m and 1/100m was used for θ, since spatial effects on a much smaller scale than this would be impossible to detect because of uncertainty about exact home locations and spatial effects on a much larger scale than this would appear as large-scale trends. Thus in our model the location- and attribute-specific risk intensity per hundred population is given by

$$\Lambda(x) = \frac{e^{X_0(x)\beta_0}}{(1-\bar{X}_0) + \bar{X}_0\, e^{\beta_0}} \left[\sum_{j=1}^{4} X_j(x)\beta_j + \frac{\theta^2}{2\pi} \int_{\mathcal{S}} e^{-\theta^2|y-s|^2/2}\, \Gamma(ds) \right]. \quad (4)$$

The latent term $X_5(x)\beta_5 = (\theta^2/2\pi)\int_{\mathcal{S}} e^{-\theta^2|y-s|^2/2}\,\Gamma(ds)$ may be thought of as a location-dependent or spatial random effect, while the observed term $X_4(x)\beta_4$ (NO_2) can be viewed as a location-dependent fixed effect, from the usual Generalized Linear Mixed Model perspective. This model generalizes the ones described in Ickstadt and Wolpert (1999) through the inclusion of a multiplicative relative risk factor. Indeed, with the exception of the spatial random effect, it is possible for all risk factors to enter multiplicatively in the spirit of the more familiar relative risk models used in epidemiology. Our present approach offers the flexibility to include both multiplicative and additive risk factors as appropriate, although the choice of which form to assume will usually be subjective.

Figure 3 presents the structure of our random field model visually in the form of a conditional independence graph using the notation introduced by Spiegelhalter (1998) for Bayesian graphical models. Ovals represent random variables or random measures in the model, boxes represent known quantities (observed covariates, etc.) and arrows indicate direct dependencies: solid lines for stochastic relationships, broken lines for deterministic ones. The stack of unshaded boxes (indexed by $j = 0, ..., 4$ in our example) represents the observed covariates and attributes; they need not be conditionally independent given y, as the figure suggests. The small shaded box represents the latent spatial covariate ($j = 5$). The large shaded box expresses how these influence in turn first the Poisson mean density $\Lambda(x)$, then the mean measure $\Lambda(dx)$, and finally the (observed) point counts $N(dx)$. Prior distributions for the parameters β_j and θ are not shown explicitly. See Appendix A for a discussion of the computations.

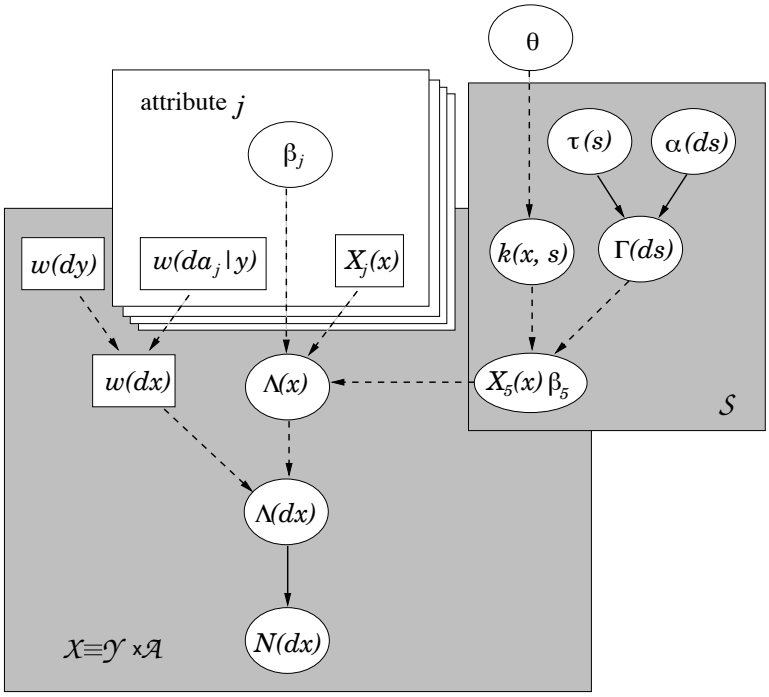


Figure 3. Graphical representation of the random field model.

6. RESULTS

Table 1 reports the posterior mean and 90% credible interval for the relative risk of severe wheeze for boys compared to girls, given by the posterior distribution of $\exp(\beta_0)$. Table 2 reports the posterior mean and 90% credible interval for the number of cases per 100 children aged seven to nine years that are attributed to each of the excess risk factors baseline (j=1), dampness (j=2), passive smoking (j=3), NO_2 exposure (j=4) and the latent spatial factor (j=5). These are given by the posterior distribution of $\bar{X}_j\beta_j = \int_{\mathcal{X}} X_j(x)\beta_j\, w(dx)/w(\mathcal{X})$ for $j = 1, \ldots, 4$, and $\bar{X}_5\beta_5 = \int_Y X_5(y)\beta_5\, w(dy)/w(\mathcal{Y})$ (recall that in this application, $X_5(x) = X_5(y)$). Also reported in Table 2 are the posterior means $\bar{\beta}_j$ of the regression parameters for $j = 1, \ldots, 4$: $\bar{\beta}_1$ is just the baseline rate of severe wheeze per 100 children; $\bar{\beta}_2$ and $\bar{\beta}_3$ may be interpreted as the number of additional cases of severe wheeze expected per 100 children exposed to home dampness or passive smoking respectively compared to 100 unexposed children; $\bar{\beta}_4$ is the additional number of cases per 100 children expected for every $\mu g/m^3$ increase in elevated NO_2. Estimates for the latter suggest that in an area with annual mean NO_2 concentration 41 $\mu g/m^3$ above the nominal baseline (the maximum concentration observed in Huddersfield), we would expect the rate of severe wheeze to increase by $0.01 \times 41 = 0.41$ cases per 100 children. Corresponding estimates of all parameters for the model without the latent spatial covariate $X_5(x)$ are also given for comparison.

Together the baseline and (when present) latent spatial covariate account for more than three-quarters of the risk. The inclusion of a latent spatial covariate decreases the attribution to each of the other contributing factors and increases the uncertainty (i.e., widens the interval estimates) for most of them, especially the baseline risk and the relative risk for boys to girls. Others have noted increased uncertainty for fixed effects when random effects are included in regression modeling (Rodriguez and Goldman 1995; Carlin 1999). The posterior distribution for relative risk of boys to girls in the presence of the latent spatial covariate is highly skewed (the median is only 1.85:1, nearly the same as without the latent spatial covariate), while the log relative risk is not.

The posterior (5% trimmed) mean of the latent spatial covariate $X_5(x)\beta_5 = \theta^2/(2\pi) \int_S e^{(-\theta^2|y-s|^2/2)}\Gamma(ds)$ is shown in Figure 4, through shading proportional to the unattributed disease risk per 100 children aged seven to nine years, averaged over (a) the polygonal EDs, (b) an arbitrary grid of approximately 1 km squares and (c) continuously. Our random field model allows

Table 1. Posterior summaries for the relative risk parameter for identity link models with and without a latent spatial covariate.

	With Latent Spatial Covariate		**Without Latent Spatial Covariate**	
	Relative Risk		*Relative Risk*	
Risk Factor	*Mean*	*90% Int.*	*Mean*	*90% Int.*
Boys : Girls	2.96 : 1	[.49, 8.03]	1.85 : 1	[1.44, 2.33]

Table 2. Posterior summaries for each excess risk parameter for identity link models with and without a latent spatial covariate; results are reported for the posterior mean coefficient $\bar{\beta}_j$ (for $j = 1, ..., 4$), the posterior mean and 90% interval for the number of cases attributed to each excess risk factor per 100 children at risk, and the posterior mean percentage of cases attributed to each excess risk factor.

	With Latent Spatial Covariate			**Without Latent Spatial Covariate**		
	$\bar{\beta}_j$	*No. (%) cases per 100*		$\bar{\beta}_j$	*No. (%) cases per 100*	
Risk Factor		*Mean*	*90% Int.*		*Mean*	*90% Int.*
Dampness	.69	.18= 8.1%	[.01, .41]	1.07	.28=11.6%	[.11, .42]
Smoking	.28	.08= 3.5%	[.01, .24]	.40	.11= 4.5%	[.01, .25]
NO_2	.01	.10= 4.4%	[.01, .39]	.02	.16= 6.9%	[.01, .51]
Baseline	.28	.28=12.8%	[.01, 1.65]	.18	1.82=77.0%	[1.45, 2.09]
Latent covariate		1.57=71.2%	[.083, 2.187]			
Base+ Latent		1.85=84.1%	[1.31, 2.22]		1.82=77.0%	[1.45, 2.09]

risk associated with spatial covariates to be presented consistently at multiple spatial scales. Figures 4(d) and (e) show the lower and upper quartile of this posterior distribution averaged over an arbitrary grid, respectively.

Figures 4 (a)–(c) shows evidence of local, small scale variation in unattributed risk but little or no broad spatial trend. The rate of unattributed cases per hundred is close to the overall average of 1.57 over most of the map, suggesting only weak residual spatial dependence. This is further supported by the posterior concentration near zero of the inverse spatial scale parameter θ, with $\mathsf{E}[\theta^{-1}] \approx 14.5\,\text{km}$ comparable to the $28.2\,\text{km} \times 19.6\,\text{km}$ dimensions of

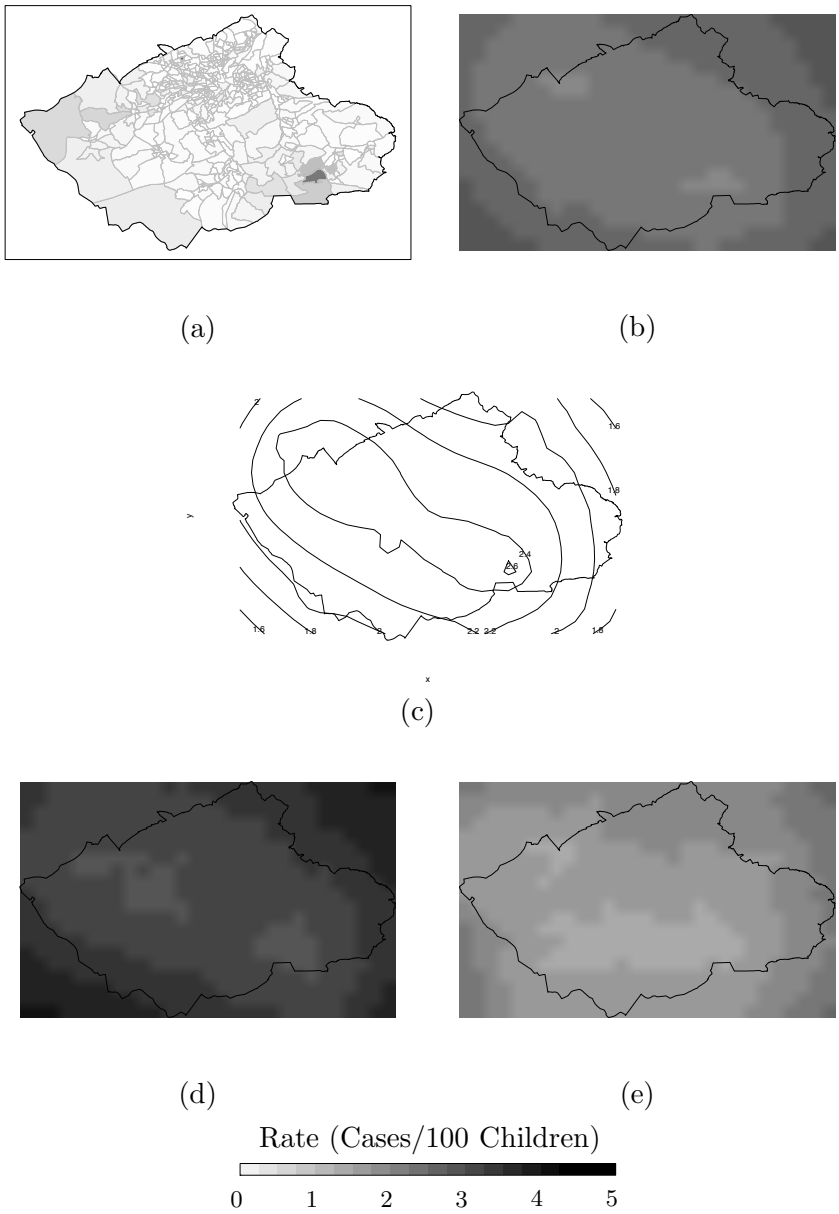


Figure 4. Posterior expected rate (in cases/100 children) attributed to the latent covariate (a) in each enumeration district, (b) on an arbitrary grid, and (c) continuously. Maps (d) and (e) show, respectively, the lower and upper quartiles of this posterior distribution on the arbitrary grid.

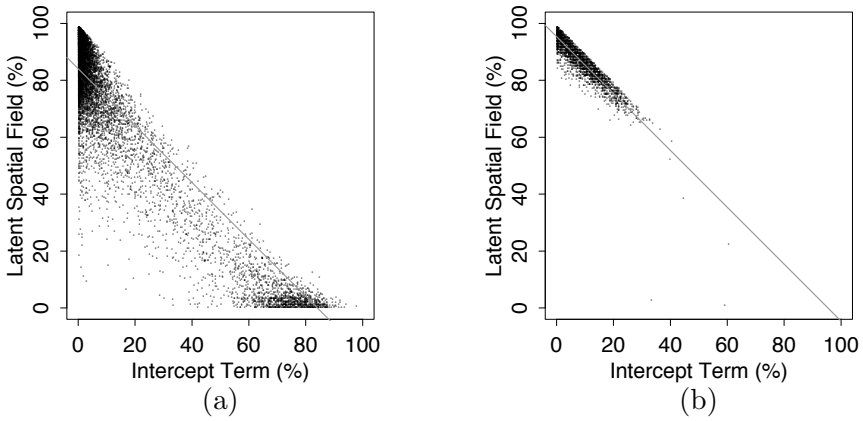
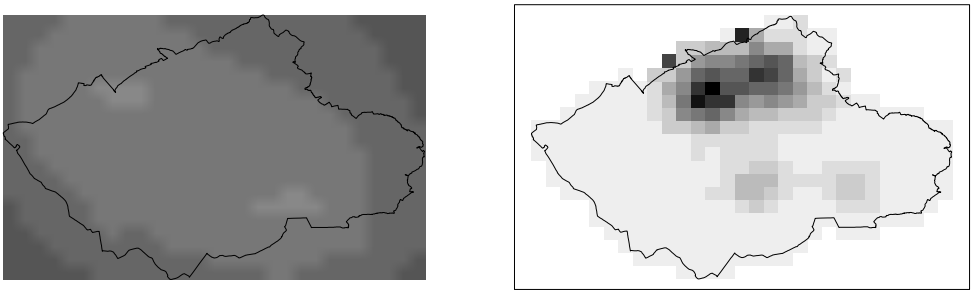


Figure 5. Bivariate distribution of posterior risk attribution (in percent) associated with the latent covariate effect $X_5(x)\beta_5$ and the baseline β_1, for the models with (a) population-based and (b) area-based reference measure. Diagonal line is best-fit with slope -1.

Huddersfield. This suggests that any "spatial" effects occur at such a large spatial scale that they are nearly independent of location within our study region. There is one small area in the south-east, apparent on each of the ED, grid and contour presentations (a)–(c), where residual rates appear to be somewhat elevated. There is also an area of slightly raised residual risk in the north west of Huddersfield, more apparent on the grid presentation than the ED or contour maps. The quartile plots (d) and (e) suggest that there is some uncertainty about the overall magnitude of the latent risk; the relative high spots in the south-east and north-west of Huddersfield persist however.

Both the baseline term β_1 and the latent spatial covariate $X_5(x)\beta_5$ reflect unattributed risk. If the risk does *not* feature spatial variation, the two terms will be confounded by multicollinearity. The joint posterior distribution of β_1 and β_5 presented in Figure 5(a) (scaled as percent of total unattributed risk) suggests that indeed there is collinearity, offering further evidence of the absence of residual spatial variation. This leads to wide marginal posterior distributions for the effects attributed to the baseline and latent spatial factor but a well-determined posterior value for their sum, as was apparent from Table 2.

To investigate whether separation of the baseline and latent spatial factor is achieved when the unattributed risk *does* feature strong spatial variation, we repeated our analysis using an area-based uniform population reference



(a) (b)

Figure 6. Posterior expected total risk attributed to the latent covariate, for the models with (a) population-based and (b) area-based reference measure.

measure $w_{\mathrm{Y}}(dy) = \bar{w}\,dy$ based on the average population density $\bar{w}$ of seven to nine year old children across the whole of Huddersfield. This gives $\Lambda(x) = \Lambda(y, a)$ the interpretation of a conditional disease rate *per unit area* for children with attribute value a at location y, and we would expect the latent covariate $X_5(x)$ to capture the spatial variation in population density. Figure 5(b) shows that the joint posterior for β_1 and $X_5(x)\beta_5$, while still collinear, is now much more concentrated on the latent spatial field. Figure 6 contrasts the posterior mean of the latent spatial covariate with (a) population-weighted reference measure and (b) area-weighted reference measure, and illustrates that without population-based weighting the model would misattribute risk to a latent spatial covariate concentrated in the high-population areas. Under area weighting the posterior mean inverse precision is $\mathsf{E}[\theta^{-1}] \approx 1.85\,\mathrm{km}$, again reflecting much stronger spatial dependence.

7. DISCUSSION AND FURTHER RESEARCH

We believe this to be the first presentation of a spatial epidemiological model supporting coherent inference from both individual attributes and from spatial evidence given on different, non-nested spatial scales. Others have considered spatial evidence given on non-nested partitions (e.g., Mugglin, Carlin, and Gelfand (1998)), but the spatial aggregation required in that approach would appear to preclude associating individual attributes and risk factors with cases.

The latent portion of the disease intensity in our approach may be viewed as a collection of infinitely many unobserved point sources of uncertain strength. Most of the point source literature employs *pre-specified* point sources (see Diggle (1990), or Lawson and Waller (1996) and references therein), but recently some have considered latent point source models (Lawson and Clark 1999), an approach that could be extended using marked point processes to support regression on individual attributes as well as spatial covariates.

We now consider how well the present approach overcomes the inherent limitations of earlier approaches to studying the SAVIAH data we discussed in Section 3. This analysis makes full use of the available covariates to support inference, and in particular, allows individual level covariates to be treated as case-specific even when other data enter the model in aggregate form. While it is capable of reflecting the variability of children's NO_2 exposure at home and at school, the analysis we have presented still employs only a point estimate of home NO_2 concentration and so does not overcome that limitation (see Section 7.2 for further discussion). We do however allow for the effect of a spatially correlated but unmeasured additive excess risk factor. Our approach avoids aggregating data to coarser geographical units, and so does not introduce additional areal distortion and increase ecological bias. However we cannot undo any aggregation inherent in the data we have and so there is still some possibility for ecological bias. Our NO_2 data were available only on a $250\,\mathrm{m} \times 250\,\mathrm{m}$ grid, for example, with unavoidable implicit extrapolation errors; our case locations were available only at the postcode level, and cannot express fully the variability of exposure which the subjects experience.

An important feature of our models is the flexibility they offer to allow both excess and relative risk factors to be included in the same model. Admittedly, the choice between additive or multiplicative forms for the covariate effects is largely subjective and should be made in consultation with subject matter experts. The additive component of the model is also limited to 'harmful' effects (i.e., which are associated with increased risk of disease). However, there are usually *a priori* scientific grounds to argue that certain covariate effects act in a specific direction. For example, there are no plausible mechanisms to support the hypothesis that damp walls, maternal smoking or ambient air pollution genuinely protect against wheezing illness, and so a model which specifies a non-negative relationship for these risk factors seems reasonable. The flexibility of our model may be further enhanced by introducing interaction terms, although the difficulty

of implementing formal model choice criteria in the present framework (see Section 7.2 for further discussion) suggests that such an approach must be guided by scientific considerations and should be viewed as a form of sensitivity analysis.

7.1 Epidemiological Implications

As in many other epidemiological studies of childhood respiratory illnesses, no single risk factor emerged as dominant in our analysis. Less than one quarter of the total risk was attributed to specific sources and, of these, home dampness was found to account for nearly twice as much risk of severe wheeze as did either passive smoking or NO_2. We also found that boys experience approximately 1.5–3.0 times higher risk than girls. These are in accord with earlier analyses of the SAVIAH data by Fischer et al. (1998) and with analyses of a different data set by Strachan and Carey (1995).

None of the previous studies of the association of childhood wheeze and asthma with exposure to road traffic pollution summarized by Elliott et al. (1997) found a significant association of elevated risk of respiratory illness with interpolated NO_2 exposure. However, over half of those studies that used either traffic flow measurements or distances to roads as proxies for exposure to traffic pollution did report small but significant associations with wheeze or asthma.

Consistent with earlier studies of traffic-related pollution we found a small elevation of risk associated with elevated levels of NO_2 exposure, accounting for perhaps 4–5% of cases. It is possible that the effect is in fact greater, but is masked by unmodeled variability of the the NO_2 covariate (from measurement or extrapolation errors, for example). We also found some suggestion of higher unexplained risk in small regions in the south-east and north-west, a matter not considered in earlier analyses. Both these regions have relatively low population but contain a number of cases (see Figure 1(a)), leading to high *observed* risk in each area. Our model should smooth these raw rates to eliminate spurious fluctuations due to sparse data. However, the model is sufficiently flexible to allow spatial variation in this smoothed risk surface which can capture localized peaks and troughs in the 'true' underlying disease rates. Investigation of the two regions highlighted by our model revealed that both are located near to the intersection of major roads (south-east region) and motorways (north-west region). Therefore one possibility is that the NO_2 estimates used in this analysis did not fully capture the effects of traffic-related pollution in these areas. Other

unmeasured pollution sources or risk factors may also be located nearby. These areas may therefore warrant further follow-up to identify the possible reasons for elevated rates of wheezing illness.

7.2 Model Variations and Extensions

In our analysis we have treated traffic pollution exposure as an individual attribute, based only on the NO_2 concentration at each child's home location, but children are also exposed at their schools, in their neighborhoods, and while traveling from one place to another. This variable exposure may be accommodated within the existing model framework in a variety of ways.

We could include the average NO_2 concentration at each child's recorded school location as an additional regression term $X_j(x)\beta_j$, or could replace the existing excess home NO_2 concentration X_4 with a weighted average of the concentrations at home and at school, or perhaps with a weighted average of the concentrations at locations near the home. Our results using such a weighted average (we used an isotropic Gaussian weighting centered at a nominal radius of 250 ± 250m), with or without reflecting in the model our uncertainty about the radius over which we should average NO_2 concentrations, were virtually identical to those presented here treating NO_2 as an individual attribute. Uncertainty and measurement error in the recorded NO_2 concentrations could also be accommodated by treating NO_2 as a spatial random effect, i.e., a second latent spatially varying random field whose prior mean is the reported NO_2 concentration and whose variability reflects the magnitude of the anticipated measurement errors and the degree of our uncertainty. In this application *measurement error* has several aspects: errors (or simply variability) in the original NO_2 measurements, extrapolation error (or additional and unmodeled spatial variability), exposure variability (due to children's movement within and outside their neighborhoods), and nonlinearity (the measurements only try to present time-averaged concentrations at each point, so if the effect of exposure on severe wheeze is nonlinear, perhaps depending more on the peaks than the average, then such a summary cannot be sufficient). Although we recognize the importance of measurement error for this application, full treatment of this complex issue is beyond the scope of the present analysis.

NO_2 exposure at school could also be accommodated by treating each child's location as a point in a *four*-dimensional space, two dimensions for the home location and two for school, and associating with each home/school pair an NO_2 level that reflects exposure at home, at school, and while commuting;

see Ickstadt and Wolpert (1999) for a use of four-dimensional modeling in another spatial application.

Based on earlier findings of Fischer et al. (1998) we did not include socio-economic status as a covariate in our analysis. However, measures of socio-economic deprivation are known to be important predictors of many diseases. In other UK applications it would be possible to include, for example, the Carstairs deprivation score (Carstairs and Morris 1991) based on census variables defined at the ED level, as a location dependent multiplicative relative (or additive excess) risk factor. We also omitted any possible interactions among the risk factors we consider; for example, we have treated the multiplicative relative risk factor X_0 (indicating gender) as acting uniformly on all additive excess risk factors, rather than allowing gender to act differently for different additive excess risks by including separate interaction terms $X_0(x)X_j(x)\beta_{0j}$.

Model selection and validation are notoriously difficult for complex hierarchical models. The direct computation from MCMC streams of marginal probabilities to compute Bayes factors is difficult even in parametric models, and is particularly difficult here in the presence of augmentation variables and random fields; we are exploring methods of accomplishing this, but they are beyond the scope of the present study. A possible step in this direction would be to base a residual analysis on the Deviance Information Criterion (DIC) of Spiegelhalter, Best, and Carlin (1998b).

In some applications, important individual level attributes are simply unavailable for reasons of confidentiality. In Best et al. (2000) we present an area level version of the present models that can still be used to overcome the problem of data available at incommensurate spatial resolutions.

The program code (in ISO C) is available on request from the authors. They are working in collaboration with D. Spiegelhalter and A. Thomas to explore the extension of the WinBUGS Bayesian MCMC software package (Spiegelhalter, Thomas, and Best 1998a) to include these spatial point-process regression models, to make them more widely available and easy to use.

APPENDIX A: COMPUTATIONS

The hierarchical model of Section 4. may be represented in the form:

$$
\begin{array}{lrcl}
\text{Regression} & \beta_0 & \sim & \mathsf{No}(0,1) \\
\text{Coefficients:} & \beta_j & \sim & \mathsf{Ga}(\alpha_j,\tau_j) \quad (1 \le j \le 4) \\
 & & & \alpha_j = .575, \quad \tau_j = \alpha_j \bar{X}_j \\
\text{Intensity:} & \Lambda(x) & \equiv & \dfrac{e^{X_0(x)\beta_0}}{(1-\bar{X}_0) + \bar{X}_0\, e^{\beta_0}} \displaystyle\sum_{j=1}^{4} X_j(x)\beta_j \\
\text{Point counts:} & N(dx)|_\beta & \overset{\text{ind}}{\sim} & \mathsf{Po}(\Lambda(x)\, w(dx)).
\end{array}
$$

We use Markov chain Monte Carlo (MCMC) methods to draw samples from the posterior distribution of the parameter vector $\beta = \{\beta_j\} \in \mathbb{R}^5$ by constructing an ergodic Markov chain $\beta^{(t)}$ with the posterior distribution for a stationary distribution (Tierney 1994). We must use a Metropolis-Hastings step for β_0, but by selecting at each time step t a possible risk factor to associate with each case (i.e., by augmenting the data at each step with an indicator $j_n^{(t)} \in \{1,2,3,4\}$ for each case x_n with $\mathsf{P}[j_n^{(t)} = j] \propto X_j(x_n)\, \beta_j^{(t)}$) we achieve conjugacy for the $\beta_j^{(t)}$. This conjugacy enables us to use a Gibbs sampling step for $\beta_j^{(t)} \sim \mathsf{Ga}(\alpha_j^{(t)}, \tau_j^{(t)})$ with $\alpha_j^{(t)} \equiv \alpha_j + n_j^{(t)}$ the sum of α_j and the number $n_j^{(t)} \equiv \#\{n : j_n^{(t)} = j\}$ of data points associated at time step t through our data augmentation with risk factor j, and with $\tau_j^{(t)} \equiv \tau_j + \bar{X}_j w(\mathcal{X})$ the sum of τ_j and $\int X_j(x)\, w(dx)$.

During each time step we also replace any missing covariates $X_j(x_n)$ with draws from their empirical distribution in the SAVIAH cohort.

In Section 5. we introduce a θ-dependent kernel function $k(x,s)$, a Gamma random field $\Gamma(ds) \sim \mathsf{Ga}\big(\alpha(ds), \tau(s)\big)$ with shape measure $\alpha(ds)$ and inverse scale function $\tau(s)$ on a set $\mathcal{S}$, and through them the latent spatially varying quantity $X_5(x)\beta_5 \equiv \int_{\mathcal{S}} k(x,s)\Gamma(ds)$, and expand the Poisson mean function to $\Lambda(x) \equiv \frac{e^{X_0(x)\beta_0}}{(1-X_0)+X_0\, e^{\beta_0}} \sum_{j=1}^{5} X_j(x)\beta_j$. To accommodate this in the MCMC scheme we must add at each time step a Metropolis step in which we draw a new $\theta^{(t)}$ and a Gibbs step in which we draw an entire random field $\Gamma^{(t)}(ds) \sim \mathsf{Ga}\big(\alpha^{(t)}(ds), \tau^{(t)}(s)\big)$, where again we use data augmentation to achieve a conjugate conditional distribution for the Gamma random field $\Gamma^{(t)}(ds)$; an algorithm for sampling from such random fields is presented in Appendix B. Periodically we generate and store for later display estimates of the function $\Lambda(x)$ by averaging spatially over elements of one or more selected partitions, using Monte Carlo integration for arbitrary partitions or exact Gaussian integration over rectangular elements of grid partitions. From these we compute averages, for plots of the posterior mean, or quartiles, for Figures 4(d) and (e).

APPENDIX B: GAMMA RANDOM FIELD MODELS

A random field is said to be a Gamma random field $\Gamma(ds) \sim \mathsf{Ga}\big(\alpha(ds), \tau(s)\big)$ with shape measure $\alpha(ds)$ and inverse scale function $\tau(s)$ over some set $\mathcal{S}$ if the random variables $\Gamma(A)$, $A \subset \mathcal{S}$, A measurable, follow a Gamma distribution with shape parameter $\alpha(A) = \int_A \alpha(s)ds$ and mean $\alpha(A)/\tau(a)$ for any $a \in A$ (the distribution is exact if $\tau(a)$ is constant on A, and approximate otherwise) and if the random variables $\Gamma(A)$ and $\Gamma(B)$ for sets A, B that do not overlap are independent.

The realizations of Gamma random fields are almost surely discrete, consisting of countably infinitely many point masses of magnitudes γ_i at locations s_i, enabling us to determine $\Gamma(A)$ by summing up the jumps falling into A, i.e. $\Gamma(A) = \sum_i \{\gamma_i : s_i \in A\}$.

A Dirichlet random field is the same as a Gamma random field normalized to have unit mass one, i.e. $\Gamma(ds)/\Gamma(\mathcal{S})$.

For our MCMC scheme (see Appendix A) we need to be able to generate Gamma random fields. The following sampling algorithm makes use of the fact that such a field can be constructed from a standard Poisson process and is given in detail in Wolpert and Ickstadt (1998). Firstly, generate jump places s_j uniformly on $\mathcal{S}$. Secondly, generate successive jump times t_j of a standard Poisson process. Thirdly, set $\gamma_j \equiv E_1^{-1}(t_j/\alpha(s_j))\tau(s_j)^{-1}$, where E_1^{-1} denotes the inverse of the exponential integral function $E_1(t) \equiv \int_t^\infty e^{-u}u^{-1}du$. The collection of jump places s_j and corresponding magnitudes γ_j are then a realization of the Gamma random field $\Gamma(ds) \sim \mathsf{Ga}\big(\alpha(ds), \tau(s)\big)$.

The latent term $X_5(x)\beta_5 = \int_{\mathcal{S}} k(x,s)\Gamma(ds)$ can now be calculated by

$$X_5(x)\beta_5 = \sum k(x, s_i)\,\gamma_i \tag{5}$$

using the jump places s_j and corresponding magnitudes γ_j for the realization of the Gamma random field. Thus, $X_5(x)\beta_5$ can be viewed as a smoothed version of our Gamma random field realization. For our model we use a two-dimensional Gaussian smoothing kernel, but any nonnegative integral kernel $k(x,s)$ satisfying an integrability condition to ensure that $X_5(x)\beta_5$ is well defined, might be used.

Bibliography

Bernardinelli, L., Pascutto, C., Best, N. G., and Gilks, W. R. (1997), "Disease mapping with errors in covariates," *Statistics in Medicine*, 16, 741–752.

Besag, J., York, J., and Mollié, A. (1991), "Bayesian image restoration, with two applications in spatial statistics (with discussion)," *Annals of the Institute of Statistical Mathematics*, 43, 1–59.

Best, N. G., Ickstadt, K., Wolpert, R. L., and Briggs, D. J. (2000), "Combining models of health and exposure data: the SAVIAH study," in *Spatial Epidemiology: Methods and Applications*, eds. P. Elliott, J. C. Wakefield, N. G. Best, and D. J. Briggs, Oxford: Oxford University Press, chapter 22.

Breslow, N. E. and Day, N. E. (1980), *Statistical methods in cancer research. Volume 1 - The analysis of case-control studies*, number 32 in IARC Scientific Publications, Lyon: International Agency for Research on Cancer.

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 Geographic Information Science*, 11, 699–718.

Brunekreef, B., Janssen, N. A. H., de Hartog, J., Harssema, H., and Knape, M. (1997), "Air pollution from truck traffic and lung function in children living near motorways," *Epidemiology*, 8, 298–303.

Carlin, B. P. (1999), "Discussion of 'Bayesian models for spatially correlated disease and exposure data' by Best et al." in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, pp. 131–156.

Carstairs, V. and Morris, R. (1991), *Deprivation and health in Scotland*, Aberdeen: Aberdeen University Press.

Clayton, D. G. and Kaldor, J. (1987), "Empirical Bayes estimates of age–standardized relative risks for use in disease mapping," *Biometrika*, 43, 671–681.

Committee of Medical Effects of Air Pollution (1995a), *Asthma and outdoor air pollution*, London: Her Majesty's Stationary Office.

— (1995b), *Non-biological particles and health*, London: Her Majesty's Stationary Office.

Cressie, N. A. C. and Chan, N. H. (1989), "Spatial modeling of regional variables," *Journal of the American Statistical Association*, 84, 393–401.

Diggle, P. J. (1990), "A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point," *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 153, 349–362.

Duhme, H., Weiland, S., Keil, U., Kraemer, B., Schmid, M., Stender, M., and Chambless, L. (1996), "The association between self-reported symptoms of asthma and allergic rhinitis and self-reported traffic-density on street of residence in adolescents," *Epidemiology*, 7, 578–582.

Edwards, J., Walters, S., and Griffiths, R. (1994), "Hospital admissions for asthma in preschool children: relationship to major roads in Birmingham, United Kingdom," *Archives of Environmental Health*, 49, 223–227.

Elliott, P., Briggs, D. J., Lebret, E., Goryinski, P., and Bohumil, K. (1997), "Small area variations in air quality and health (SAVIAH study): Relationship of childhood respiratory symptoms to road traffic pollution," Technical report, Imperial College, Department of Epidemiology and Public Health, Imperial College School of Medicine at St Mary's, London, UK.

Fischer, P., Kriz, B., Martuzzi, M., Wojtyniak, B., Lebret, E., van Reeuwijk, H., Pykhart, H., Briggs, D., Gorynski, P., and Elliott, P. (1998), "Prevalence and risk factors of childhood respiratory health in four countries in Western and Central Europe in the study on Small Area Variations In Air quality and Health (SAVIAH)," *Indoor Air*, 8, 244–254.

Florey, C. and Leeder, S. (1973), *Methods for cohort studies of chronic airflow linitation*, Copenhagen: WHO Regional Publications, European series No. 12.

Fotheringham, A. S. and Wong, D. W. S. (1991), "The modifiable areal unit problem in multivariate statistical analysis," *Environment and Planning*, 23, 1025–1045.

Greenland, S. (1992), "Divergent biases in Ecologic and Individual-Level Studies," *Statistics in Medicine*, 11, 1209–1223.

Her Majesty's Government and Department of Health and Department of the Environment (1996), *The United Kingdom National Environmental Health Action Plan*, London: Her Majesty's Stationary Office.

Ickstadt, K. and Wolpert, R. L. (1999), "Spatial regression for marked point processes (with discussion)," in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, pp. 323–341.

Lawson, A. B. and Clark, J. (1999), "Markov chain Monte Carlo methods for putative sources of hazard and general clustering," in *Disease Mapping and Risk Assessment for Public Health*, eds. A. Lawson, A. Biggeri, D. Boehning, E. Lessaffre, , J. Viel, and R. Bertollini, Chichester: John Wiley & Sons, pp. 119–142.

Lawson, A. B. and Waller, L. A. (1996), "A review of point pattern methods for spatial modelling of events around sources of pollution," *Environmetrics*, 7, 471–487.

Livingstone, A. E., Shaddick, G., Grundy, C., and Elliott, P. (1996), "Do people living near inner city main roads have more asthma needing treatment? Case-control study," *British Medical Journal*, 312, 676–677.

Martuzzi, M. and Elliott, P. (1996), "Empirical Bayes estimation of small area prevalence of non-rare conditions," *Statistics in Medicine*, 15, 1867–1873.

Morgenstern, H. (1998), "Ecologic Studies," in *Modern Epidemiology*, eds. K. J. Rothman and S. Greenland, Philadelphia: Lippincott, Williams and Wilkins, pp. 459–480.

Mugglin, A., Carlin, B. P., and Gelfand, A. E. (1998), "Fully model based approaches for spatially misaligned data," Technical Report 98-014, Division of Biostatsistics, University of Minnesota, to appear in *Journal of the American Statistical Association*.

Murakami, M., Ono, M., and Tamura, K. (1990), "Health problems of residents along heavy-traffic roads," *Journal of Human Ergology*, 19, 101–106.

Openshaw, S. (1984), *The modifiable areal unit problem*, volume 38 of *Concepts and techniques in modern geography*, Norwich, England: Geo Books.

Osterlee, A., Drijver, M., Lebret, E., and Brunekreef, B. (1996), "Chronic respiratory symptoms in children and adults living along streets with high traffic density," *Occupational Environment Medicine*, 53, 241–247.

Pearce, N., Weiland, S., Keil, U., Langridge, P., Anderson, H., Strachan, D., Bauman, A., Young, L., Gluyas, P., Ruffin, D., Crane, J., and Beasley, R. (1993), "Self-reported prevalence of asthma symptoms in children in Australia, England, Germany and New Zealand: an international comparison using the ISAAC protocol," *European Respiratory Journal*, 6, 1455–1461.

Pershagen, G., Rylander, E., Norberg, S., Eriksson, M., and Nordell, S. (1995), "Air pollution involving nitrogen dioxide exposure and wheezing bronchitis in children," *International Journal of Epidemiology*, 24, 1147–1153.

Pope, C., Thun, M., Namboodiri, M., Dockery, D., Evans, J., Speizer, F., and Heath, C. (1995), "Particulate air pollution as a predictor of mortality in a prospective study of US adults," *American Journal of Respiratory Critical Care Medicine*, 151, 669–674.

Richardson, S. (1992), "Statistical methods for geographical correlation studies," in *Geographical and environmental epidemiology: Methods for small-area studies*, eds. P. Elliott, J. Cuzick, D. English, and R. Stern, Oxford: Oxford University Press, chapter 17, pp. 181–204, paperback edition published 1996.

Robinson, W. S. (1950), "Ecological correlations and the behaviour of individuals," *American Sociological Review*, 15, 351–357.

Rodriguez, G. and Goldman, N. (1995), "An assessment of estimation procedures for multilevel models with binary responses," *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 158, 73–89.

Schwartz, J. (1994), "Air pollution and daily mortality: a review and meta-analysis," *Environmental Research*, 64, 36–52.

Shaw, R. A., Crane, J., Pearce, N., Burgess, C. D., Bremner, A., Woodman, K., and Beasley, R. (1992), "Comparison of a video questionnaire with the IUATLD written questionnaire for measuring asthma prevalence," *Clinical and Experimental Allergy*, 22, 561–568.

Spiegelhalter, D., Thomas, A., and Best, N. (1998a), *WinBUGS: Bayesian inference Using Gibbs Sampling, Manual Version 1.2.*, Imperial College, London and Medical Research Council Biostatistics Unit, Cambridge, available from http://www.mrc-bsu.cam.ac.uk/bugs.

Spiegelhalter, D. J. (1998), "Bayesian graphical modelling: a case-study in monitoring health outcomes," *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 47, 115–133.

Spiegelhalter, D. J., Best, N. G., and Carlin, B. P. (1998b), "Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models," Submitted to *Journal of the Royal Statistical Society, Series B (Methodological)*.

Strachan, D. and Carey, I. (1995), "Home environment and severe asthma in adolescence: A population based case-control study," *British Medical Journal*, 311, 1053–1056.

Tierney, L. (1994), "Markov chains for exploring posterior distributions (with discussion)," *Annals of Statistics*, 22, 1701–1762.

Waldron, G., Pottle, B., and Dod, J. (1995), "Asthma and motorways — one district's experience," *Journal of Public Health Medicine*, 17, 85–89.

Waller, L. A., Carlin, B. P., Xia, H., and Gelfand, A. E. (1997), "Hierarchical spatio-temporal mapping of disease rates," *Journal of the American Statistical Association*, 92, 607–617.

Wardlaw, A. J. (1992), "The role of air pollution in asthma," *Journal of Allergy and Clinical Immunology*, 90, 358–363.

Weiland, S., Mundt, K., Ruckmann, A., and Keil, U. (1994), "Self-reported wheezing and allergic rhinitis in children and traffic density on street of residence," *Annals of Epidemiology*, 4, 243–247.

Wilkinson, P., Elliott, P., Grundy, C., Shaddick, G., Thakrar, B., Walls, P., and Falconer, S. (2000), "Case-control study of hospital admissions with asthma in children aged 5–14 years: relation with road traffic in North West London," *Thorax*, (in press).

Wjst, M., Reitmeir, P., Dold, S., Wulff, A., Nicolai, T., Freifrau von Loeffelholz-Colberg, E., and von Mutius, E. (1993), "Road traffic and adverse effects on respiratory health in children," *British Medical Journal*, 307, 596–600.

Wolpert, R. L. and Ickstadt, K. (1998), "Poisson/gamma random field models for spatial statistics," *Biometrika*, 85, 251–267.