A Critique of Ecological Studies

Jonathan Wakefield

NRCSE

Technical Report Series

NRCSE-TRS No. 072

November 13, 2001

The NRCSE was established in 1997 through a cooperative agreement with the United States Environmental Protection Agency which provides the Center's primary funding.

A Critique of Ecological Studies

Jonathan Wakefield

Departments of Statistics and Biostatistics, University of Washington, Seattle, USA,

Department of Epidemiology and Public Health, Imperial College School of Medicine,
London, UK

In many ecological studies investigating associations between environmental exposures and health outcomes, the observed relative risks are in the range 1.0–2.0. The interpretation of such small relative risks is difficult due to a variety of biases, some of which are unique to ecological studies. In this paper the mathematical assumptions of a number of commonly-used ecological regression models are made explicit, critically assessed, and related to ecological bias. Methods for determining the likely effects of unmeasured confounding and within-area variability in exposures/confounders are also described. These methods may also be useful for deciding on the utility of an ecological study, by explicitly considering the likely size of the association, the strength of confounders and the extent of within-area variability in exposures/confounders. In a well-designed ecological study in which known confounders are collected, it is argued that an observed association can only be deemed plausible if the strength of the association is 'large', or if within-area individual-level data on exposures and confounders are incorporated. With regard to the latter we discuss two-phase designs that would appear to be useful in environmental settings. The modelling of spatial variability is considered and related to the underlying continuous spatial field. The examination of such a field with respect to the modelling of risk in relation to a point source highlights an inconsistency of commonly-used approaches. It is argued that the sophistication of the statistical analysis should not outweigh the quality of the data.

Keywords: confounding; ecological fallacy; exposure misclassification; pure specification bias; spatial epidemiology.

1 Introduction

The aim of this paper is to discuss a number of issues relating to ecological studies in the context of environmental epidemiology. Such studies utilize data at the level of the group rather than the individual, and have a long history in epidemiology (Morgenstern, 1998), as well as sociology (Robinson, 1950), political science (Cleave, Brown and Payne, 1995, King, 1997) and geography (Openshaw, 1984). In this paper we concentrate on non-infectious rare diseases.

Ecological studies are controversial due to a variety of difficulties that are summarized under the umbrella term of *ecological bias*. Incorrect conclusions may be reached in ecological studies for a variety of reasons (Greenland, 1992) including: *pure specification bias* which arises when a nonlinear individual exposure/risk model is assumed to apply at the area level, within- and between-area *confounding*, *errors-in-variables*, *effect modification* and the lack of *mutual standardisation* (Rosenbaum and Rubin, 1984). Confounding and exposure misclassification are not unique to ecological studies but can also lead to incorrect inference in an individual level study. There is a vast literature on ecological bias and the ecological fallacy in the epidemiological literature, see for example, Greenland and Morgenstern (1982), Piantodosi, Byar and Green (1988), Richardson, Stucker and Hemon (1987), Greenland and Robins (1994).

Ecological studies have a variety of aims including: investigation of the variability of relative risks across a region (disease mapping), examining the association between risk and environmental exposures that may be in air, water or soil (ecological regression), surveillance of routine health statistics for early detection of 'hot spots' of risk (cluster detection), and investigation of risk in relation to a putative pollution source (point or line source studies). Recent statistical developments in environmental epidemiology may be found in the edited volumes by Lawson et al. (1999) and Elliott et al. (2000).

In studies of environmental pollution from point sources in developed countries, unless there is an accident resulting in a large increase of pollutant (such as that at Chernobyl), the increases in risk are often modest. Occupational studies tend to produce much larger increases. A number of point source studies have been carried out by the Small Area Health Statistics Unit (Elliott et al. 1992b) in the UK. Examples include: all incinerators of waste solvents and oils in Great Britain (Elliott et al. 1992a); a single petrochemical works at Baglan Bay, Wales (Sans et al. 1995); radio and TV transmitters (Dolk et al. 1997a,b); municipal incinerators (Elliott et al. 1996), cokeworks (Dolk et al. 1999); a pesticides factory (Wilkinson et al. 1997); landfill sites (Dolk et al. 1998); and industrial complexes that include major oil refineries (Wilkinson et al. 1999). These have reported excesses in risk at source in the range 0.1–1.0 that is, relative risks of 1.1–2.0, and are consistent with the value of 1.5 quoted by Pekkanen and Pearce (2001) as being typical. These relative risks must be viewed in light of the fact that for cancers in particular there are risk factors that are far more predictive of disease than environmental factors, for example diet, smoking, alcohol consumption and genetic factors. Consequently the potential for confounding is strong since ecological studies do not directly utilise individual-level risk factor data (though stratification by age, gender and socio-economic status is routinely carried out). In Elliott et al. (1992b) a point source study was carried out to investigate increased risk of meseothelioma in the vicinity of Plymouth docks. This analysis revealed an estimated excess of 11 at source, but further analysis revealed that this excess was due to occupational, rather than environmental, risk factors. In the context of occupational epidemiological studies Siemiatycki et al. (1988) investigated the biases that occur in estimates of relative risk when the variables smoking, ethnic group and socio-economic status are not incorporated in the analysis. Their conclusions were that for lung cancer relative risks in excess of 1.4 are unlikely to be artifacts due to uncontrolled confounding while for bladder and stomach cancer the equivalent figure was 1.2. These figures should be viewed as a lower bound of acceptability for ecological studies since, as noted above, the within-area variability in exposures/confounders leads to the potential for a variety of other biases. The effects of the bias not only cast doubt on the conclusions of studies that reveal a small detremental effect, but also on studies that reveal no association.

The problems of interpretation in ecological study arise in large part from data quality issues, beyond the observational and aggregate nature of such studies. In particular, because of the routinely-collected nature of much of the data utilized in environmental epidemiology errors in numerator, denominator and in exposure variables may be extremely influential. For cancer cases and hospital admissions, retrospectively, registry (Best and Wakefield, 1999) and provider (Aylin et al. 2001) effects may also be significant. With respect to the denominators, the effects of under-enumeration at census and migration must be considered. These and other data anomalies are not likely to be spatially neutral and could be a significant contribution to observed variation in risk estimates. Such errors have been described in detail by Elliott and Wakefield (1999) and will not be concentrated upon here though the acknowledgement of their presence is vital, both when statistical analyses are interpreted, and when the level of statistical sophistication of the analysis is considered. In particular, for rare diseases analysed at the small area level a small number of cases may effectively drive the analysis and the sensitivity of the conclusions should be assessed by removing these influential cases.

The structure of this paper is follows. In Section 2 we introduce notation and review a number of approaches that are currently used to analyse ecological data. Section 3 considers how sensitivity to unmeasured confounding may be carried out in ecological studies, a topic that has received little attention. Such sensitivity analyses have a long history in epidemiology, see Rothman and Greenland (1998, Chapter 19) for references. In Section 4 the bias due to aggregation of a nonlinear exposure/risk model is considered and a number of proposed solutions are critically reviewed. The results of Section 3 are applied in order to produce a means by which the effects of within-area variability in exposures/confounders may be assessed. A Bayesian non-parametric model for modelling within-area variability is also proposed. Section 5 provides an interpretation of random effects models that are often utilised in ecological analyses. In Section 6 the modeling of a continuous risk surface is considered. In the context of the modelling of disease risk in relation to a point/line source of pollution a number of inconsistencies in parametric approaches are highlighted. In Section 7 we discuss design issues, and in particular how the utility of a particular study may be determined. Section 8 contains a concluding discussion.

2 Statistical Framework

2.1 Conventional Approaches

We consider a study area A that may be partitioned into sub-areas A_i, $i = 1, ..., n$, according to data availability. Within area i we suppose there are $N_i = \sum_{c=1}^{C} N_{ic}$ individuals where N_{ic} denotes the number of individuals in confounder stratum c, $c = 1, ..., C$. Typically these stata will represent age and gender, and possibly a measure of socio-economic status.

Richardson and Montfort (2000) provide a review of approaches to ecological inference in this context. A basic model for a rare disease is to assume

$$Y_i | R_i \sim \text{Poisson}(E_i \times R_i), \tag{1}$$

where $E_i = \sum_j N_{ic} p_c$ with p_c a 'reference probability', and R_i is the *relative risk* of area i. In Section 6 we discuss the exact interpretation of R_i. A simple approach to ecological inference is to obtain the MLE $\hat{R}_i = Y_i/E_i$, and regress $\log \hat{R}_i$ on μ_i^x, an area-level measure of exposure via an additive or multiplicative model. Cook and Pocock (1983) carry out an analysis within this framework, accounting for spatial dependence via a simple spatial error model.

The model (1) suffers from a number of difficulties. The MLE $\hat{R}_i$ is well-known to be highly variable for rare diseases and so high/low values for particular areas may just be a reflection of sampling variability. Also, environmental epidemiological data often display overdispersion, that is the variance exceeds the mean. As described in Section 5, this variability may be due to unmeasured risk factors with or without spatial dependence. Both imprecision due to small numbers, and overdispersion, may be addressed via the introduction of random effects. Besag, York and Mollié (1991) proposed the model

$$\log R_i = \beta_0 + \beta_1 \mu_i^x + T_i + S_i, \tag{2}$$

where $T_i|\sigma_t^2 \sim N(0, \sigma_t^2)$ denote unstructured (independent) random effects, and S_i random effects with spatial structure. Besag, York and Mollié (1991) modelled the latter using the intrinsic conditional autoregressive (ICAR) model in which one considers the conditional distribution of $S_i|S_j, j \in \partial i$ where ∂i represents the indices of a set of 'neighbouring' areas. Specifically, $S_i|S_j, j \in \partial i \sim N(\bar{S}_i, \sigma_s^2/m_i)$, where $\bar{S}_i = \frac{1}{m_i}\sum_{j\in\partial i} S_j$ and m_i denotes the number of neighbours. This model has the advantage of being computationally straightforward to implement and non-stationary. The latter is appealing in terms of flexibility, though the level of non-stationarity, in terms of the set of risks that may be represented has not been investigated. For example, it is unlikely that discontinuities due to geographical features such as rivers or mountains could be well-modelled with a normal model, though a double exponential model may be more amenable to such features (Besag, York and Mollie, 1991; Best et al. 1999). If discontinuities of this type are expected then appropriate covariates may be incorporatd in (2). The model does not consider the positions, sizes and shapes of the areas, and the interpretation of σ_s^2 is difficult unless each area has a constant number of neighbours, which makes prior elicitation and interpretation more difficult. The non-stationarity also makes model checking more troublesome since the joint model is improper and so there is no distribution with which marginal random effect estimates may be compared (though the specification of one random effect, or the mean of the collection, yields a proper prior).

An alternative approach is to model the joint collection $S = (S_1, ..., S_n)'$. An example that has been considered by, amongst others, Best et al. (1999) and Wakefield and Morris (1999), is $S|\Sigma_s \sim N_n(0_n, \Sigma_s)$ where 0_n denotes an $n \times 1$ vector of zeroes, and $\Sigma_{sij} = \sigma_s^2 \exp(-d_{ij}\phi)$ where d_{ij} is the distance between the centroids of areas i and j. This model has the advantage of simple interpretation of the parameters σ_s^2 and ϕ, so that prior distributions are more straightforward to specify; σ_s^2 is a marginal variance and so is directly comparable to σ_t^2 and $\log 2/\phi$ denotes the distance at which correlations fall to 0.5. We note that measuring the extent of spatial variability as a function of the total variability is not straightforward since ϕ needs to be considered also; one possibility is to consider the posterior distribution of $|\Sigma_s|$. The joint model also ignores the topology of the areas and is computationally intensive. The stationarity may be restrictive, and the importance of the choice of correlation function also needs to be determined. For both the conditional and the joint models, realisations of T_i, S_i and $\log R_i$ may be generated and examined to assess the level of residual variability, spatial and otherwise.

Rather than model conditionally or jointly, it is more natural to construct a model from the underlying continuous risk surface, an approach we discuss in Section 6.

2.2 Individual-Level Models

We now describe a hypothetical individual-level model. We let $Y_{icj} = 0/1$ represent the event that individual j in stratum c of area i is a non-case/case. We denote exposures of interest by X_{icj}, and confounders by U_{icj}. We then have the *individual*-level model

$$\mathrm{E}[Y_{icj}|X_{ki}, U_{icj}] = p(X_{icj}, U_{icj}). \tag{3}$$

Then

$$Y_{icj}|X_{icj}, U_{icj} \sim \text{Bernoulli}\{p(X_{icj}, U_{icj})\}.$$

We may have an *additive model*

$$p(X_{icj}, U_{icj}) = \beta_0 + X_{icj}\beta_1 + U_{icj}\beta_2, \tag{4}$$

in which β_1, β_2 represent *risk differences*, or a *multiplicative model*

$$p(X_{icj}, U_{icj}) = \exp(\beta_0 + X_{icj}\beta_1 + U_{icj}\beta_2), \tag{5}$$

where e^{β_1}, e^{β_2} are *relative risks*. Here and throughout we assume there are no contextual effects so we do not, for example, consider models of the form

$$p(X_{icj}, U_{icj}) = \beta_0 + X_{icj}\beta_1^w + \bar{X}_i\beta_1^b + U_{icj}\beta_2.$$

Sheppard (2001) contains discussion of such models. We have also assumed that the risk parameters are constant across areas.

3 Sensitivity to an Unmeasured Confounder

In this section we discuss a number of models that may be used to assess the potential effect of unmeasured confounding, both between and within areas. Such sensitivity studies have a long history in epidemiology, beginning with Cornfield et al. (1959). In this section, for clarity, we assume no measured confounders and consider various scenarios.

3.1 Additive Model

The most general model we consider is given by

$$E[Y_{ij}|X_{ij}, U_{ij}] = \beta_0 + \beta_1 X_{ij} + \beta_2 U_{ij}. \tag{6}$$

In the following let $\mu_i^x = E[X_{ij}|i]$ and $\mu_i^u = E[U_{ij}|i]$. If there is no between-area confounding (so that μ_i^x and μ_i^u are independent) then

$$E[Y_{ij}|\mu_i^x] = \beta_0^* + \beta_1\mu_i^x,$$

where $\beta_0^* = \beta_0 + \beta_2\mu_i^u$, and so, even if there is within-area confounding, no bias will result. It also follows that regardless of the within-area behaviour, if we have measured all confounders at the area-level, there will be no confounding. If an individual-level study were carried out in any one area, however and U were unmeasured then bias would result, showing that ecological studies can provide improvements on individual-level studies. One of the examples of Greenland and Robins (1994) is based on this scenario.

3.1.1 Binary variables

Suppose X and U are both binary in which case

$$\mu_i^x = \Pr(X_{ij} = 1|i) \quad \text{and} \quad \mu_i^u = \Pr(U_{ij} = 1|i).$$

We define $\Pr(U_{ij} = 1|X_{ij} = x, i) = P_{ix}$, $x = 0, 1$, $i = 1, ..., n$, so that $\mu_i^u = \Pr(U_{ij} = 1|i) = \Pr(U_{ij} = 1|\mu_i^x = 1, i) = P_{i0} + (P_{i1} - P_{i0})\mu_i^x$. Under this model we have confounding both within and between areas. No confounding corresponds to $P_{i0} = P_{i1} = P_i$. In general we have

$$E[Y_{ij}|\mu_i^x] = \beta_0 + \beta_2 P_{i0} + \{\beta_1 + \beta_2(P_{i1} - P_{i0})\},$$

and so when the model

$$E[Y_{ij}|\mu_i^x] = \beta_{0i}^\star + \beta_{1i}^\star \mu_i^x, \tag{7}$$

is fitted we have

$$\beta_{0i}^\star = \beta_0 + \beta_2 P_{i0}$$

and

$$\beta_{1i}^\star = \beta_1 + \beta_2(P_{i1} - P_{i0}).$$

As an example of sensitivity to within-area confounding, suppose that $Y = 0/1$ represents absence/presence of lung cancer, $X = 0/1$ low radon/high radon and $U = 0/1$ affluent/deprived, so that $P_{ix} = \Pr(\text{deprived}|\text{radon } x, i)$. Suppose P_0 =0.1, P_1=0.3 so that the probability of being deprived is three times higher if resident in a high radon area. If there is no effect of radon on lung cancer ($\beta_1 = 0$) and β_2 =0.1, but we do not measure deprivation then we would estimate the risk difference as $\beta_1^\star = 0.02$. Conversely if there is an effect of β_1 =0.02, this could be lost if β_2 =0.05, P_0 =0.5, P_1 =0.1 (to give $\beta_1^\star = 0$).

We now briefly consider the situation in which we have an interaction, i.e.

$$E[Y_{ij}|X_{ij}, U_{ij}] = \beta_0 + \beta_1 X_{ij} + \beta_2 U_{ij} + \beta_3 X_{ij} U_{ij}.$$

We have

$$E[X_{ij}U_{ij}|\mu_i^x, \mu_i^u] = P_{i1}\mu_i^x,$$

which allows the sensitivity to be addressed. Laserre et al. (1999) examine this model when only μ_i^x and μ_i^u are measured and advocate

$$E[X_{ij}U_{ij}|\mu_i^x, \mu_i^u] \approx \mu_i^x \mu_i^u$$

which corresponds to independence, that is, $P_{i1} = \mu_i^u$.

3.1.2 Normal variables

Now consider the case of no interaction and continuous exposure/confounder. Again we assume that an area-level summary of the exposure is available only. Recall that when an additive model is appropriate, we only need to worry about between-area confounding. A convenient between-area model is given by

$$\begin{bmatrix} \mu_i^x \\ \mu_i^u \end{bmatrix} \sim N\left(\begin{bmatrix} \mu^x \\ \mu^u \end{bmatrix}, \begin{bmatrix} \Sigma^x & \Sigma^{xu} \\ \Sigma^{ux} & \Sigma^u \end{bmatrix} \right), \tag{8}$$

where $\Sigma^{ux} = \rho(\Sigma^x \Sigma^u)^{1/2}$.

With model (6), if we regress on μ_i^x only we obtain (7) with

$$\begin{aligned} \beta_0^\star &= \beta_0 + \beta_2\{\mu^u - \mu^x(\Sigma^u/\Sigma^x)^{1/2}\rho\}, \\ \beta_1^\star &= \beta_1 + \beta_2(\Sigma^u/\Sigma^x)^{1/2}\rho. \end{aligned} \tag{9}$$

As expected the effect will be overestimated if $\beta_2 > 0$ and $\rho > 0$. The extent of the bias is determined by the ratio of the standard deviations of the confounder to the exposure which is intuitively reasonable.

Water constituents such as magnesium and calcium, or magnesium and a continuous measure of socio-economic status, are examples of continuous exposures/confounders for which this model may be useful. For example if $\Sigma^x = \Sigma^u$ then $\beta_1^\star$ =0.05 could be obtained from $\beta_1 = 0$, ρ =0.5 and β_2 =0.1. This model may simply extended to multiple confounders, as described in the next section.

3.2 Multiplicative Model

Now suppose we have the model

$$E[Y_{ij}|X_{ij}, U_{ij}] = \exp(\beta_0 + \beta_1 X_{ij} + \beta_2 U_{ij}). \tag{10}$$

In the absence of between-area confounding, within-area confounding may lead to bias, in contrast to the additive model case. If we wanted to simulate a scenario in which there were within-area confounding but no between-area confounding we could assume μ_i^x and μ_i^u were independent and then model the within-area confounding via odds ratios ψ_i, $i = 1, ..., n$.

3.2.1 Binary variables

We first suppose that X_{ij} and U_{ij} are binary and as with the additive model assume $P_{ix} = \Pr(U = 1|\mu_i^x = x, i)$. We will examine the bias that results when the model

$$E[Y_{ij}|\mu_i^x] = \exp(\beta_0^* + \beta_1^* \mu_i^x), \tag{11}$$

is fitted, which is equivalent to

$$E[Y_{ij}|\mu_i^x] = \alpha_0^* + \alpha_1^* \mu_i^x, \tag{12}$$

where $\alpha_0^* = \exp(\beta_0^*)$ and $\alpha_1^* = \exp(\beta_0^*)\{\exp(\beta_1^*) - 1\}$ and the relative risk $\exp(\beta_1^*) = 1 + \alpha_1^*/\alpha_0^*$.

A loglinear regression of Y_{ij} on μ_i^x produces

$$E[Y_{ij}|\mu_i^x] = \exp(\beta_{i0}^* + \beta_{i1}^* \mu_i^x),$$

where

$$\begin{aligned} \beta_{0i}^\star &= \beta_0 + \log[(1 - P_{i0}) + P_{i0}\exp(\beta_2)], \\ \beta_{1i}^\star &= \beta_1 + \log\left\{\frac{(1 - P_{i1}) + P_{i1}\exp(\beta_2)}{(1 - P_{i0}) + P_{i0}\exp(\beta_2)}\right\}. \end{aligned}$$

Lin, Psaty and Kronmal (1999) obtain this form when $\mu_i^x = 0/1$ is constant within areas. An obvious way to examine sensitivity to between-area confounding is to assume that $P_{i0} = P_0$ and $P_{i1} = P_1$. For such a choice, an observed relative risk of $e^{\beta_1^\star}$ =1.2 could be obtained with $\beta_1 = 0$, e^{β_2} =1.5, P_0 =0.1 and P_1 =0.5. The chance of not measuring a confounder with a relative risk of 1.5 that is five times more prevalent in exposed than non-exposed individuals may be viewed as unlikely, but with multiple confounders the relative risks and strength of dependence are reduced (as is demonstrated in the next section).

3.2.2 Normal variables

Now consider continuous exposure/confounders and suppose that (8) applies. If we assume that there is no within-area variability in exposures or confounders then coefficients of model (11) are given by (9). We stress that here we are interested in bias, the distribution of the data is no longer Poisson in general, and in particular the variance is larger than the mean.

The extension to multiple confounders is straightforward. Suppose

$$E[Y_i|X_i, U_{1i}, ..., U_{Ci}] = \exp\left(\beta_0 + \beta_1 X_i + \beta_2 \sum_{c=1}^{C} U_{ci}\right).$$

Then under multivariate normality of $X_i, U_{1i}, ..., U_{Ci}$ with $\text{corr}(X_i, U_{ci}) = \rho$ and $\text{var}(X) = \text{var}(U_{ci})$ we have

$$E[Y_i|X_i] = \beta_0^\star + \beta_1^\star X_i,$$

where

$$\beta_1^\star = \beta_1 + \beta_2 C \rho,$$

so that if $\beta_1 = 0$ we have $\beta_1^\star = \beta_2 \times C \times \rho$ showing the exact interplay between number of confounders, strength of dependence, and resultant estimate (when there is no association). For example $C = 4$ confounders each having a correlation of 0.25 with the exposure, and e^{β_2} =1.2 yield an observed relative risk of $e^{\beta_1^\star}$ =1.2.

Now consider the situation in which we have within and between area confounding but we measure both μ_i^x and μ_i^u. If we assume

$$\begin{bmatrix} X_{ij} \\ U_{ij} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_i^x \\ \mu_i^u \end{bmatrix}, \begin{bmatrix} \Sigma_i^x & \Sigma_i^{xu} \\ \Sigma_i^{ux} & \Sigma_i^u \end{bmatrix}\right),$$

then we obtain

$$E[Y_{ij}|\mu_i^x, \mu_i^u] = \exp\{\beta_1 \mu_i^x + \beta_2 \mu_i^u + (\beta_1^2 \Sigma_i^x + \beta_2^2 \Sigma_i^u + 2\beta_1\beta_2\Sigma_i^{xu})/2\}. \quad (13)$$

The terms $\exp(\beta_1^2 \Sigma_i^x/2)$ and $\exp(\beta_2^2 \Sigma_i^u/2)$ arise due to pure specification bias (see next section) while the within-area confounding is responsible for the term $\exp(2\beta_1\beta_2\Sigma_i^{xu}/2)$. If the variances/covariances are constant, or independent of μ_i^x then no bias will result.

3.3 Overdispersion

In practice, one indicator of the presence of unmeasured variables is the extent of overdispersion. Quasi-likelihood provides a simple method for estimating the latter via specification of the first two moments only. Alternatively, random effects models, such as those described in Section 2.1 may be fitted. For the models considered here a natural choice is

$$\mathrm{var}(Y_i|\mu_i^x) = E[Y_i|\mu_i^x](1 + E[Y_i|\mu_i^x] \times c). \tag{14}$$

One situation that leads to this relationship occurs when we have

$$E[Y_i|\mu_i^x, \mu_i^u] = \exp(\beta_0 + \beta_1\mu_i^x + \beta_2\mu_i^u),$$

with μ_i^x and μ_i^u independent, in which case

$$E[Y_i|\mu_i^x] = E_i \exp(\beta_0^* + \beta_1\mu_i^x). \tag{15}$$

If $\delta_i = \exp(\beta_2\mu_i^u) \sim_{iid} \mathrm{Ga}(a^u, b^u)$ (so that the unmeasured variables follow a log gamma distribution) we have (15) with $\beta_0^* = \beta_0 + \log a^u - \log b^u$, and (14) with $c = 1/a^u$. The choice $\mu_i^u \sim_{iid} N(\mu^u, \Sigma^u)$ gives $\beta_0^* = \beta_0 + \beta_2\mu^u + \beta_2^2\Sigma^u/2$ and $c = \exp(\Sigma^u) - 1$. This latter model will be considered in Section 5. The specification

$$\delta_i = \exp(\beta_2\mu_i^u) \sim \mathrm{Ga}\{\ E_i \exp(\beta_0 + \beta_1\mu_i^x)b^u E[\exp(\beta_2\mu_i^u)],\ \ E_i \exp(\beta_0 + \beta_1\mu_i^x)b^u\ \}$$

leads to (15) with $\beta_0^* = \beta_0 + \log E[\exp(\beta_2\mu_i^u]$ and $\mathrm{var}(Y_i|\mu_i^x) = E[Y_i|\mu_i^x](1+\kappa)$ with $\kappa = 1/b^u$. This variance function was assumed by Diggle et al. (1997) as a pragmatic means of incorporating extra-Poisson variability. This model is not so easily-interpretable, however, since

$$\mathrm{var}(\delta_i) = \frac{E[\exp(\beta_2\mu_i^u)]}{E_i \exp(\beta_0 + \beta_1\mu_i^x)b^u}.$$

Although the above models do not examine bias due to confounding, a large value of c or κ does indicate there are unmeasured variables (or data anomalies), some of which may be confounders, and indicate that caution should be exercised when interpreting observed associations.

4 Pure Specification Bias

In this section we assume for simplicity that there are no confounders and consider in isolation the effect of aggregation of the individual exposure/risk model, Greenland (1992) has termed this *pure specification bias*.

4.1 Parametric Approach

We assume a univariate continuous exposure and that within area i, $X|\phi_i \sim f(\cdot|\phi_i)$ where ϕ_i denotes a set of parameters that characterise the exposure. Then, for individual j in area i, $j = 1, ..., n_i$,

$$Y_{ij}|\beta, \phi_i \sim \mathrm{Bernoulli}\{p(\phi_i)\},$$

where

$$p(\phi_i) = \mathrm{E}_{X|\phi_i}[p(X)] = \int p(x)f(x|\phi_i)\mathrm{d}x.$$

If exposures are independent within areas, and the outcome is rare

$$Y_i|\beta,\phi_i \sim \text{Poisson}\{n_i p(\phi_i)\},$$

where $Y_i = \sum_{j=1}^{n_i} Y_{ij}$.

If we have the additive model $p(x) = \beta_0 + \beta_1 x$ then no bias arises since $p(\phi_i) = \beta_0 + \beta_1\mu_i^x$ where $\mu_i^x = \text{E}[X|i]$. For the multiplicative model $p(x) = \exp(\beta_0 + \beta_1 x)$ we have

$$p(\phi_i) = \exp(\beta_0)\text{E}[\exp(X\beta_1)] \tag{16}$$

(Richardson, Stucker and Hemon 1987) where the expectation is with respect to $X|\phi_i$. For the case $X|\phi_i \sim N(\mu_i^x, \Sigma_i^x)$ with $\phi_i = (\mu_i^x, \Sigma_i^x)$ where $\mu_i^x = E[X_{ij}|i]$ and $\Sigma_i^x = \text{var}(X_{ij}|i)$ we obtain

$$p(\phi_i) = \exp(\beta_0 + \beta_1\mu_i^x + \beta_1^2\Sigma_i^x/2). \tag{17}$$

In the unlikely event that Σ_i^x is constant across areas there will no bias (Plummer and Clayton, 1996). We describe a simple method for determining the extent of the bias using ideas from the last section. Suppose that across areas we have

$$\Sigma_i^x \approx a + b\mu_i^x,$$

then

$$\text{E}[Y_{ij}|\mu_i^x] \approx \exp(\beta_0^\star + \beta_1^\star\mu_i^x),$$

with $\beta_0^\star = \beta_0 + a\beta_1^2/2$, $\beta_1^\star = \beta_1 + b\beta_1^2/2$. Hence if $\beta_1 > 0$ and, as we would expect, the variance increases with the mean then ignoring within-area variability will lead to overestimation of β_1. If $\beta_1 < 0$ so that the exposure is protective, then with $b > 0$ the size of the effect will be underestimated and may even change sign. Hence in the magnesium study reported by Maheswaran (1999), although no protective effect of magnesium was found, this could have been lost due to within-area variability in magnesium (which was substantial, Wakefield and Morris, 1999). It has been recognised that many ecological studies find larger effects than their individual-level counterparts.

Many environmental exposures are well-modelled by lognormal distributions but the moment generating function of a lognormal does not exist and so (16) cannot be evaluated. Wakefield and Salway (2001) give the form of this function for a gamma within-area distribution.

As a final example we consider the case in which the within-area distribution is approximated by a uniform distribution. This may provide a method of assessing the sensitivity when in each area some measure of the spread in exposure is available. The choice $X|\phi_i \sim U(\mu_i^x - c_i, \mu_i^x + c_i)$, with $\phi_i = (\mu_i^x, c_i)$, gives

$$p(\phi_i) = \exp(\beta_0 + \beta_1\mu_i^x)\frac{(e^{\beta_1 c_i} - e^{-\beta_1 c_i})}{2c_i\beta_1},$$

for $\beta_1 \neq 0$. Hence there is no bias if c_i is constant across areas. Greenland (1992) also considered this choice and examined via simulation the effect on estimation of a uniformly distributed exposure and a uniformly distributed covariate that were independent across areas.

There are a number of disadvantages to the parametric approach. In particular the distribution $X|\phi_i$ needs to be known and sufficient within-area samples are required for accurate estimation of ϕ_i. Wakefield and Salway (2001) show how for small within-area samples the estimation of the variance in particular is highly unstable and can lead to inaccurate inference. They also illustrate how the parametric approach may account for the measurement error model and this approach was used to model the relationship between childhood leukaemia and benzene by Best et al. (2001). To alleviate the instability, one possibility is to model the variance as a smooth function of the mean.

4.2 Aggregate Data Approach

Prentice and Sheppard (1995) consider the situation in which survey data on individual exposures and confounders is available on a subset of m_i $(2 \leq m_i \leq n_i)$ individuals from each area, see also Sheppard and Prentice (1995).

The likelihood is analytically intractable but the first two moments may be readily evaluated and so an estimating functions approach is possible where each of the mean, variance and derivatives can be approximated using the subsample $X_i^{m_i}$. For example the mean is approximated by $m_i p(X_i^{m_i})$ where

$$p(X_i^{m_i}) = \mathrm{E}[Y_{ij}|X_i^{m_i}] = \frac{e^{\beta_0}}{m_i} \sum_{j=1}^{m_i} \exp(X_{ij}^T \beta_1). \tag{18}$$

Each of the area survey averages for the mean, variance and vector of derivatives are unbiased with respect to the sampling distribution of $X_i^{m_i}$ but for any specific realisation, bias will be present in the estimating function containing these estimates. Prentice and Sheppard (1995) discuss this bias in the case of $V_k(\beta) = 1$ and determine an unbiased estimating function by explicitly calculating the bias.

There are a number of advantages to the aggregate data approach and many of the problems with the standard ecological analysis can be overcome, see Guthrie and Sheppard (1999) and Wakefield and Salway (2001).

If a closed form expression cannot be found for a particular within area exposure distribution for area k, $f(\cdot|\phi_i)$, then one may proceed using Monte Carlo integration. Specifically we may estimate $\mathrm{E}[\exp(X^T \beta_1)]$ by

$$\hat{p}_1(\phi_i) = \frac{e^{\beta_0}}{M_i} \sum_{j=1}^{M_i} \exp(X_j^T \beta_1), \tag{19}$$

where $X_j|\phi_i \sim_{iid} f(\cdot|\phi_i)$, $j = 1, ..., M_i$ represents a random sample. A similar approach may be taken for more general risk/exposure models.

4.3 A Bayesian Non-Parametric Approach

In the previous sections we have described approaches for dealing with within-area variability based on the assumption of a parametric distribution for this variability and on surveys of sampled values. It is still an open question as to the size of such surveys that are required but it seems clear that very small surveys will provide very little information for reliable inference (and this was seen in the limited simulations of Wakefield and Salway, 2001). In this section we will describe a speculative procedure that is intended for those situations in which we have sparse survey data but additional information, for example some idea of the mean and variance of the within-area exposure distribution.

The approach is Bayesian and is based on the Dirichlet process prior of Ferguson (1973). We assume that the univariate exposure within area i, X_{ij} arise from the unknown distribution function F_i. From a Bayesian perspective we need to place a prior distribution on this distribution function. We assume a Dirichlet process prior, denoted DPP(F_{0i}, α_i) where F_{0i} is a baseline measure that is our prior guess of the distribution and α_i may be viewed as the sample size associated with the specification F_{0i} (see below).

Now suppose we observe a sample of exposure measurements within the areas, $X_{ij}, j = 1, ..., m_i, i = 1, ..., N, m_i \geq 0$. We then have

$$\begin{aligned} p(X_i^{m_i}, F_{0i}, \alpha_i) &= \text{E}[\exp(X_i\beta_1)|X_i^{m_i}, F_{0i}, \alpha_i] \\ &= e^{\beta_0}\left\{\frac{\alpha_i}{\alpha_i + m_i}\text{E}[\exp(X_i\beta_1)|F_{0i}] + \frac{m_i}{\alpha_i + m_i}\frac{1}{m_i}\sum_{j=1}^{m_i}\exp(X_{ij}^T\beta)\right\}, \quad (20) \end{aligned}$$

a weighted combination of the prior guess and the average of the sampled exposure/risk models. Here $\text{E}[\exp(X_i^T\beta)|F_{0i}]$ is the cumulant generating function with respect to the distribution F_{0i} (the normal distribution is an obvious choice). This expression also illustrates why α_i may be viewed as a *prior sample size.* Note that $\alpha_i = 0$ produces the aggregate data approach of Section 4.2 and $m_i = 0$ the parametric approach of Section 4.1; and that, in general, some areas may have no information concerning the parametric form, or any survey information. We may use model (20) for inference from either a likelihood or Bayesian perspective.

5 Interpretation of Random Effects Models

5.1 Residual Modeling

When model (2) is fitted, one interpretation of the random effects is that they are accounting for unmeasured confounders, and for errors in the data. To be more explicit, suppose the 'true' model is given by $Y_i|R_i \sim_{iid}$ Poisson(E_iR_i), but we observe $E_i^\star$ where

$$\log E_i = \log E_i^\star + T_i, \quad (21)$$

$T_i|\sigma_t^2 \sim_{iid} N(0, \sigma_t^2)$, so that we have a Berkson errors-in-varables model (e.g., Carroll, Ruppert and Stefanski, 1995) for data anomalies. Further assume that

$$\log R_i = \beta_0 + \beta_1\mu_i^x + \beta_2\mu_i^u, \quad (22)$$

where the risk factors $\mu_i^u \sim N(\mu_u, \Sigma_u)$, $i = 1, ..., n$.

Now suppose we observe $\{Y_i, \mu_i^x, E_i^\star\}$, $i = 1, ..., n$ and assume the model $Y_i|R_i^\star \sim_{iid}$ Poisson$(E_i^\star R_i^\star)$. We then obtain

$$\log R_i^\star = \beta_0^\star + \beta_1^\star\mu_i^x + T_i + S_i,$$

with

$$\beta_0^\star = \beta_0 + \sigma_t^2/2 + \mu_u \beta_2,$$

$T_i|\sigma_t^2 \sim_{iid} N(0, \sigma_t^2)$, $(S_1, ..., S_N)'|\Sigma_s \sim N(0, \Sigma_s)$ and $\Sigma_s = \beta_2^2 \Sigma_u$, which is identical to model (2). This make clear that the random effects T_i are equal to $\mu_i^u - \mu_u$, $i = 1, ..., n$.

In the above we have assumed that the risk factors are constant within areas. Sections 3 and 4 indicate that the random effects could also be representing within-area variability in exposures/confounders; for example T_i and S_i may be soaking up the $(\beta_1^2 \Sigma_i^x + \beta_2^2 \Sigma_i^u + 2\beta_1\beta_2\Sigma_i^{xu})/2$ term in equation (13). This provides some backing to the statement of Bernardinelli et al. (1995, p. 2436) that: 'A cluster size bigger than the area size leads to a [spatial structured] *clustering* model, while a cluster size smaller than the area size leads to a *heterogeneity* model'. We note that in Section 3 we described models that were marginalised across unmeasured area-level variables, here we are considering conditional models. In the former case the response will usually be no longer Poisson (due to overdispersion) while in the latter the Poisson assumption may be reasonable.

In general $\beta_1^\star \neq \beta_1$ due to unmeasured confounding (Clayton, Bernardinelli and Montomoli, 1993), this is elaborated upon in Section 3. This issue remains one of the most difficult in environmental epidemiology and is independent of within-area variability in exposure. Even the decision as to whether to include measures of latitude and longitude in the linear predictor can have a large impact on the regression coefficient of the exposure of interest, possibly removing part of the exposure effect (if there is a north-south or east-west trend in the exposure). The decision essentially comes down to whether one believes there are unmeasured confounders with longitude or latitude trends (so that the latter are plausible surrogates).

To summarise, random effects may be thought to be accommodating any or all of data anomalies, within-area variability in risk factors, and unmeasured between-area risk factors. The spatial pattern of such quantities determines whether spatial or non-spatial random effects are dominant.

5.2 Non-Constant Relative Risks

We now consider the situation in which data from multiple point or line sources are available. Such a design is appealing since confounding is less likely over multiple sites, and problems of the lack of an a priori hypothesis are not present. A number of multiple-site studies have been carried out, for example Elliott et al. (1997) considered 72 municipal incinerators in the United Kingdom, analysing each separately via Stone's test and combining the resultant p-values, and Dolk et al. (1998) examined 21 European landfill sites and modeled the relative risks from each as random effects. We examine a number of situations that motivate such models. We assume that

$$Y_i|\theta_i \sim \text{Poisson}\{n_i \exp(\theta_i)\}, \tag{23}$$

for $i = 1, ..., N$, where $\exp(\theta_i)$ represents the relative risk associated with 'proximity' (for example, within 2km) to point/line-source i, and n_i denotes the number of individuals close to i, $i = 1, ..., N$. For clarity we have ignored stratification variables, these may be considered by replacing n_i by E_i. At the second stage of the model, the log relative risks are assumed to arise from some distribution, with the usual choice being the normal.

Again it is beneficial to begin at the level of the individual. We assume that $E[Y_{ij}|X_{ij}] = \exp(\beta_0 + \beta_1 X_{ij})$, where X_{ij} represents the exposure of individual j who is close to point source i, $i = 1, \cdots, N$, $j = 1, \cdots, n_i$.

In the first scenario suppose that $X_{ij} \sim N(\mu_i^x, \Sigma_i^x)$ denotes the exposure distribution of individuals in proximity to site i. Then we obtain (23) with $\theta_i = \beta_0 + \beta_1\mu_i^x + \beta_1^2\Sigma_i^x/2$. A constant relative risk is obtained only if the exposure mean and variance are constant across sites. Hence in this scenario the random effect is accounting for between-site variability in exposure.

In the second situation we again suppose there is a constant effect but that there is an unmeasured between-site variable μ_i^u. This leads to (23) with

$$\theta_i = \beta_0 + \log E[\exp(\beta_1 X)] + \log E[\exp(\beta_2 \mu_i^u)],$$

where the first expectation is with respect to the exposure distribution and the second expectation is with respect to the distribution across areas of μ_i^u.

Finally the randon effects model could be due to effect modification by area so that at the individual level the relative risk for site i is given by $\exp(\beta_{1i})$, $i = 1, \cdots, N$. Such modification could be due to an interaction between the effect of exposure and characteristics of the individuals in area i (for example the socio-economic status of the individuals).

The plausibility that at least one of non-constant exposure distributions/unmeasured variables/effect modification would indicate that it would be beneficial to incorporate random effects into the modelling of relative risks when multiple point/line-source data are available.

6 Continuous Risk Surface Modelling

The majority of the approaches to risk modelling with aggregate data have directly modelled the quantities R_i in equation (1). An alternative that has intuitive appeal is to consider the underyling relative risk *surface*, $R(s)$, where s denotes spatial location. A variety of choices exist for the underlying continuous model, the difficulty lies in the aggregation step. Diggle (1990) originally considered Poisson point process models for non-aggregate data, Wolpert and Ickstadt (1998) and Best, Ickstadt and Wolpert (2001) gamma random field models, and Kelsall and Wakefield (2001) Gaussian random field models. Diggle, Tawn and Moyeed (1998) discuss Gaussian random field models in the context of point data.

We begin by assuming that cases follow a Poisson process with intensity

$$\lambda(s) \times p(s),$$

where $\lambda(s)$ represents the population at risk at spatial location s. For area i we then have

$$Y_i \sim \text{Poisson}(N_i \times e^{\beta_0} \times R_i),$$

where e^{β_0} is the overall risk,

$$R_i = \int_{A_i} f(s)R(s)\text{d}s,$$

$R(s)$ is the *relative risk*, and $f(s)$ is a distribution representing the population density at location s. The naive interpretation is that R_i represents the relative risk of each of the individuals within area i. This is only true if $R(s) = R_i I(s \in A_i)$, that is, constant relative risk. The correct interpretation is that R_i is the average relative risk with respect to $f(s), s \in A_i$.

In many cases we may be able to model $f(s)$ as piecewise uniform, i.e. as

$$f(s) = I(s \in A_{ik}) \times f_{ik}, \quad k = 1, ..., K_i,$$

where K_i is the number of uniform subregions and

$$\sum_{k=1}^{K_i} f_{ik} |A_{ik}| = 1.$$

In this case we obtain

$$R_i = \sum_{k=1}^{K_i} R_{ik} f_{ik} |A_{ik}|,$$

where $R_{ik} = |A_{ik}|^{-1} \int_{A_{ik}} R(s)\mathrm{d}s$. This formulation shows that we have a deconvolution problem and to identify $R(s)$ we need to propose a model since there are an infinite number of possible collections $\{R_{i1}, ..., R_{iK_i}\}$ that result in any particular value of R_i.

As a simple example we assume that the only spatial variability in risk arises from an exposure $X(s)$ via a multiplicative exposure/risk model. In this case we have $R(s) = \exp\{\beta_1 X(s)\}$. If we sample individuals in area i according to $f(\cdot)$ and $X(s) \sim N(\mu_i^x, \Sigma_i^x)$ then we obtain

$$R_i = \exp(\beta_1 \mu_i^x + \beta_1^2 \Sigma_i^x / 2) = e^{-\beta_0} \times p(\phi_i),$$

where $\phi_i = (\mu_i^x, \Sigma_i^x)$, as in Section 4.1. The Poisson model is appropriate if the exposures are independent within areas (which is unlikely to be true for environmental exposures).

A number of authors have considered the modelling of disease risk in relation to a point source of pollution using a parametric exposure/risk model (e.g. Diggle, 1990; Diggle and Rowlingson 1994; Lawson, 1993; Diggle et al. 1997; Wakefield and Morris, 2001). Lawson (1993) considers risk as a function of orientation and allows a non-monotonic distance-risk relationship, the remainder of the approaches model risk as a monotonic function of distance. All approaches ignore the aggregate version of the model, however. Diggle and Elliott (1995) have discussed the problems of aggregation in the situation of modelling risk in relation to a point source.

We consider the model

$$R(s) = 1 + \alpha \exp(-\beta ||s - s_0||), \tag{24}$$

where s_0 is the location of a putative point source (Diggle, 1990). The ecological model

$$R_i = 1 + \alpha \exp(-\beta \delta_i^{ave}) \tag{25}$$

where δ_i^{ave} represents the population-weighted centroid is that which is often assumed. Let $\delta = ||s - s_0||$ then, from (24)

$$R_i = 1 + \alpha \int_{D_i} f(\delta) \exp(-\beta\delta) \mathrm{d}\delta,$$

where $f(\delta)$ represents the population density as a function of distance. If, for example, we assume $f(\delta)$ is uniform on $(\delta_i^{ave} - \delta_i^0, \delta_i^{ave} + \delta_i^0)$, then we obtain

$$R_i = 1 + \alpha \exp\left\{ -\beta \delta_i^{ave} \frac{(e^{\beta\delta_i^0} - e^{-\beta\delta_i^0})}{2\delta_i^0} \right\}. \qquad (26)$$

Model (26) will not necessarily be monotonic decreasing in δ_i^{ave} though areas of the same size and uniform density give the closest link between the point and ecological models. If information is available on the population density within area i then (26) may be directly used. Although it is important to be aware of the inconsistency between (24) and (25), the latter was initially proposed as a model that could pick up broad trends in the risk surface and so its use is still merited, though estimated risk/distance functions should not be over-interpreted. Similarly estimates for particular areas should be viewed with caution.

Stone (1988) proposed a test of the hypothesis $H_0 : R_1 \geq R_2 \geq ... \geq R_n$ where the areas have been ordered so that R_i represents the relative risk in the area that is $i-$th closest to s_0. Again this approach is not consistent when the underlying population density is considered since monotonic $R(\delta)$ is not equivalent to monotonic R_i.

In this paper the temporal behaviour of exposures has not been considered. We now briefly discuss a space/time formulation and assume that cumulative exposure is relevant. In general we have $Y_i \sim \text{Poisson}(N_i \times e^{\beta_0} \times R_i)$, where

$$R_i = \int_{-T}^{T} \int_{A_i} \exp\{\beta_1 X(s,t)\} \mathrm{d}s\mathrm{d}t,$$

and $(-T, T)$ is the study period. As a simple example suppose $X(s,t) \sim N(\mu_i^x + b_i^x t, \Sigma_i^x)$ so that the exposure follows a linear relationship over time (though the relationship is area-specific). Then

$$R_i = \exp(\beta_1 \mu_i^x + \beta_1^2 \Sigma_i^x / 2) \times c_i$$

where $c_i = \left[\exp(\beta_1 b_i^x T) - \exp(-\beta_1 b_i^x T)\right] / \beta_1 b_i^x$ may be ignored if b_i^x is independent of i, i.e. if time trends are constant across areas. Hence the naive ecological regression in which the relative risk is regressed on μ_i^x implicitly assumes that any within-area variability in exposure is constant across areas, and that time trends are similar across areas.

This brief discussion indicates that it is beneficial to have exposure/confounder data available both within areas and across time to gain an understanding of the spatio-temporal exposure surface. Shaddick and Wakefield (2001) carried out an analysis of daily monitored levels of four air pollutants across eight sites in order to inform a study of the health effects of acute pollution. They found that for each pollutant the majority of the variability was across time and so modelling the spatial variability was of secondary importance.

7 Ecological Study Design

There has been little consideration of design in an ecological setting, Plummer and Clayton (1996) and Sheppard, Prentice and Rossing (1996) are two exceptions. The discussion of Sections 3–6 indicate that beyond the need for exposure variability across areas and a consideration of data quality issues, there is a need to understand the likely extent of within- and between area confounding and the within-area variability in exposure (which unfortunately is likely to be larger if between-area contrasts are large).

As described in Section 4, the fundamental problem of ecological inference is the within-area variability in exposures and confounders. The collection of individual-level data is vital to help alleviate ecological bias. Richardson, Stucker and Hemon (1987) proposed making parametric assumptions on exposures and confounders within areas, while Prentice and Sheppard (1995) described an alternative strategy in which subsamples of exposures and confounders were obtained within areas. Both of these procedures require samples within areas, the latter explicitly, and the former implicitly in order to estimate the relevant moments (and to examine the appropriateness of the assumed distributional form).

Prentice and Sheppard (1995) proposed their method in the context of the situation in which samples were routinely-available through surveys for example, hence random sampling was utilized. Here we consider the situation in which an environmental epidemiological study is envisaged and individual-level data may be collected. Non-random sampling via two-phase approaches may be carried out to increase efficiency, and have proved useful in a range of epidemiological contexts (White, 1982; Breslow and Cain, 1988; Breslow and Holubkov; 1998; Scott and Wild, 1997).

Here we briefly describe one possible two-phase design that may be used in an environmental epidemiological setting. As a context we describe a study that investigated death from myocardial infarction as a function of the water constituents magnesium, calcium, fluoride and lead, with known confounders age, gender and socio-economic status (Maheswaran et al. 1999). For each area we have the number of cases and the population at risk, by age and gender. In the notation of Breslow and Chatterjee (1999) we let $S = j, j = 1, ..., J$ index the J stratum that consist of a stratification of the exposure variables and the confounders. For example, each of the N individuals in the study region can be assigned a low/high value of each of the water constituents and socio-economic status, based on the area in which they live. Along with (say) ten age bands and gender we therefore have $2^3 \times 2 \times 10 \times 2 = 320 = J$ strata. Sampling within the $2J$ stratum for cases and controls may then be carried out. Breslow and Cain (1988) advocate sampling numbers as equal as possible within each stratum for improved efficiency. A crucial assumption here is that once we have adjusted for confounders and exposures, the area is no longer important as a predictor, this will be inappropriate as will usually be the case when there is clustering of unmeasured risk factors within areas. To overcome this we are currently working on a hybrid two-stage, two-phase design in which the clustered sampling is accounted for. Korn and Graubard (1999) discuss multistage sampling in the context of health surveys. A difficulty with the overall approach is the retrospective collection of the exposure data, diseases with short latencies (congenital malformations, for example) would therefore be more amenable to studies of this type. Studies investigating the link between air pollution and hospitalisation may also be well-suited to this design.

8 Discussion

In this paper we have described a number of the sources of bias in ecological studies. Although the modelling of residual spatial variability in risk is important, it will usually be of secondary importance when compared to assessing/modelling the effect of unmeasured confounding and within-area variability. The existence and usual extent of data anomalies also indicate that robust methods are important in this setting.

The discussion of Section 5 indicates that when ecological regression analyses are carried out, both spatial and non-spatial random effects should be included, and if multiple point/line-sources are considered then a random effects approach is recommended. When mutiple sites are considered the model (23) may be considered, or (25) with α and β, the parameters describing the risk-distance relationship, allowed to be random effects. Such a model was described by Wakefield and Morris (2001) though whether the quality of the data support such a choice must be evaluated on a case-by-case basis. Whenever random effects are included addressing the sensitivity of inference to the prior distributions, particularly on variance components, is vital. As with all random effects modelling, it is also important to try to determine sources of variability; in Section 5 a number of potential sources were described.

If 'small' relative risks are envisaged then within-area samples of exposures/confounders are essential if any faith is to be placed in observed associations. In particular an understanding of the spatial and temporal variability in exposures, and the role played by measurement error is essential for choosing an appropriate statistical model.

Acknowledgements

The author would like to thank Paul Elliott for numerous conversations on the utility of ecological studies, and Ruth Salway for detailed comments on an earlier draft. SAHSU is sponsored by the Departments of Health and the Environment and the Health and Safety Executive. Although the research described in this article has been funded in part by the United States Environmental Protection Agency through agreement CR825173-01-0 to the University of Washington, it has not been subjected to the Agency's required peer review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

References

Aylin, P., Bottle, A., Wakefield, J., Jarup, L., Elliott, P. (2001). Proximity to coke works and hospital admissions for respiratory and cardiovascular disease in England and Wales. *Thorax*, **56**, 228–233.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M. and Songini, M. (1995). Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine*, **14**, 2433–2443.

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.

Best, N.G., Arnold, R.A., Thomas, A., Waller, L.A. and Conlon, E.M. (1999). Bayesian models for spatially correlated disease and exposure data. In *Bayesian Statis-*

tics 6 (J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds.), pp. 131–156, Oxford University Press, Oxford.

Best, N.G., Cockings, S., Bennett, J., Wakefield, J. and Elliott, P. (2001). Ecological regression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias. *Journal of the Royal Statistical Society, Series A*, **164**, 155–174.

Best, N.G. and Wakefield, J.C. (1999). Accounting for inaccuracies in populations counts and case registration in cancer mapping studies. *Journal of the Royal Statistical Society, Series A*, **162**, 363–382.

Best, N.G., Ickstadt, K. and Wolpert, R.L. (2001). Spatial Poisson regression for health and exposure data measured at disparate spatial scales. *Journal of the American Statistical Association*, **95**, 1076–1088.

Breslow, N.E. and Cain, K.C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, **75**, 11–20.

Breslow, N.E. and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Applied Statistics*, **48**, 457–468.

Breslow, N.E. and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase outcome-dependent sampling. *Journal of the Royal Statistical Society, Series B*, **59**, 447–461.

Carroll, R.J., Ruppert, D., and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*, Chapman and Hall.

Clayton, D., Bernardinelli, L. and Montomoli, C. (1993). Spatial correlation in ecological analysis. *International Journal of Epidemiology*, **22**, 1193–1202.

Cook, D.G. and Pocock, S.J. (1983). Multiple regression in goegraphical mortality studies, with allowance for spatially correlated errors. *Biometrics*, **39**, 361–371.

Cleave, N., Brown, P.J. and Payne, C.D. (1995). Methods for ecological inference: an evaluation. *Journal of the Royal Statistical Society, Series A*, **158**, 55–75.

Cornfield, J., Haenszel, W.H., Hammond, E.C., Lilienfeld, A.M., Shimkin, M.B. and Wynder, E.L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, **22**, 173–203.

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

Diggle, P.J. and Rowlingson, B.S. (1994). A conditional approach to point process modelling of raised incidence. *Journal of the Royal Statistical Society, Series A*, **157**, 433–440.

Diggle, P.J. and Elliott, P. (1995). Statistical issues in the analysis of disease risk near point sources using individual or spatially aggregated data. *Journal of Epidemiology and Community Health*, **49**, S20-S27.

Diggle, P.J., Morris, S.E., Elliott, P. and Shaddick, G. (1997). Regression modelling of disease risk in relation to point sources. *Journal of the Royal Statistical Society, Series A*, **160**, 491 –505.

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

Dolk, H., Elliott, P., Shaddick, G., Walls, P., and Thakrar, B. (1997a). Cancer incidence near radio and television transmitters in Great Britain: all high power transmitters, *American Journal of Epidemiology*, **145**, 10–17.

Dolk, H., Shaddick, G., Walls, P., Grundy, C., Thakrar, B., Kleinschmidt, I., and Elliott, P. (1997b). Cancer incidence near radio and television transmitters in Great Britain: Sutton Coldfield transmitter, *American Journal of Epidemiology*, **145**, 1–9.

Dolk, H., Thakrar, B., Walls, P., Landon, M., Grundy, C., Suez-Lloret, I., Wilkinson, P., and Elliott, P. (1999). Mortality among residents near cokeworks in Great Britain, *Occupational and Environmental Medicine*, **56**, 34–40.

Dolk, H., Vrijheid, M., Armstrong, B., Abramsky, L., Bianche, F., Garne, E., Nelen, V., Robert, E., Scott, J.E.S., Stone, D., and Tenconi, R. (1998), Risk of Congenital Anomalies Near Hazardous-waste Landfill Sites in Europe: The EUROHAZCON Study, *Lancet*, **352**, 423–427.

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

Elliott, P. and Wakefield, J.C. (1999). Small-area studies of environment and health. *Statistics for the Environment 4: Health and the Environment*, Barnett, V., Stein, A. and Turkman, K.F. (editors), p. 3–27, John Wiley, New York.

Elliott, P., Westlake, A., Hills, M., Kleinschmidt, I., Rodrigues, L., McGale, P., Marshall, K., and Rose, G. (1992b). The Small Area Health Statistics Unit: a national facility for investigating health around point sources of environmental pollution in the United Kingdom, *Journal of Epidemiology and Community Health*, **46**, 345–349.

Elliott, P., Shaddick, G., Kleinschmidt, I., Jolley, D., Walls, P., Beresford, J. and Grundy, C. (1996). Cancer incidence near municipal solid waste incinerators in Great Britain. *British Journal of Cancer*, **73**, 702–710.

Elliott, P., Wakefield, J.C., Best, N.G. and Briggs, D.B. (2000). *Spatial Epidemiology: Methods and Applications,* Oxford University Press, Oxford.

Ferguson, T.S. (1973). Bayesian Analysis of some non-parametric problems. *Ann. Statist.*, **1**, 209-230.

Greenland, S. (1992). Divergent biases in ecologic and individual-level studies, *Statistics in Medicine*, **11**, 1209–23.

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 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. (1999). Modelling spatial variability in disease risk. Under revision for *Journal of the American Statistical Association.*

Korn, E.L. and Graubard, B.I. (1999). *Analysis of Health Surveys*, John Wiley and Sons.

Lasserre, V., Guihenneuc-Jouyaux, C. and Richardson, S. (1999). Biases in ecological studies: utility of including within-area distribution of confounders. *Statistics in Medicine*, **19**, 45–59.

Lawson, A.B. (1993). On the analysis of mortality events associated with a prespecified fixed point. *Journal of the Royal Statistical Society, Series A*, **156**, 363–377.

Lawson, A., Biggeri, A., Bohning, D., Lesaffre, E., Viel, J.F. and Bertollini, R. (1999). *Disease Mapping and Risk Assessment for Public Health*, John Wiley.

Lin, D.Y., Psaty, B.M. and Kronmal, R.A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, **54**, 948–963.

Maheswaran, R., Morris, S., Falconer, S., Grossinho, A., Perry, I., Wakefield, J., Elliott, P. (1999). Magnesium in drinking water supplies and mortality from acute myocardial infarction in north west England. *Heart*, **82**, 455–460.

Morgenstern, H. (1998). Ecologic Study. In P. Armitage and T. Colton (Eds.), *Encyclopedia of Biostatistics Vol. 2*, pp. 1255–1276. Wiley and Sons Ltd.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem.* CATMOG No. 38, Geo Books, Norwich.

Pekkanen, J. and Pearce, N. (2001). Environmental epidemiology: challenges and opportunities. *Environmental Health Perspectives*, **109**, 1–5.

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 in ecological studies, (with discussion). *Journal of the Royal Statistical Society, Series B*, **58**, 113–126.

Prentice, R.L. and Sheppard, L. (1995). Aggregate data studies of disease risk factors. *Biometrika*, **82**, 113–25.

Richardson, S. and Montfort, C. (2000). Ecological correlation studies. In *Spatial Epidemiology: Methods and Applications.* Eds: Elliott, P., Wakefield, J.C., Best, N.G. and Briggs, D.B, pp. 205—220. Oxford University Press, Oxford.

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

Robinson, W.D. (1950). Ecological correlations and the behavior of individuals. *American Sociological Reviews*, **15**, 351–357.

Rosenbaum, P.R. and Rubin, D.B. (1984). Difficulties with regression analyses of age-adjusted rates. *Biometrics*, **40**, 437–43.

Rothman, K.J. and Greenland, S. (1998). *Modern Epidemiology, Second Edition.* Lipincott-Raven.

Sans, S., Elliott, P., Kleinschmidt, I., Shaddick, G., Pattenden, S., Walls, P., Grundy, C., and Dolk, H. (1995). Cancer incidence and mortality near the Baglan Bay petrochemical works, South Wales, *Occupational and Environmental Medicine*, **52**, 217–224.

Scott, A.J. and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, **84**, 57–71.

Sheppard, L. (2001). Insights on bias and information in group level studies. Submitted.

Sheppard, L. and Prentice, R.L. (1995). On the reliability and precision of within- and between-population estimates of relative risk parameters. *Biometrics*, **51**, 853–863.

Sheppard, L., Prentice, R.L., and Rossing, M.A. (1996). Design Considerations for estimation of exposure effects on disease risk, using aggregate data studies. *Statistics in Medicine*, **15**, 1849–1858.

Siemiatycki, J., Wacholder, S., Dewar, R., Cardis, E., Greenwood, C. and Richardson, L. (1988). Degree of confounding bias related to smoking, ethnic group, and socioeconomic status in estimates of the associations between occupation and cancer. *Journal of Occupational Medicine*, **30**, 617–625.

Shaddick, G. and Wakefield, J. (2001). Modelling multivariate pollutants at multiple sites. Under revision for *Applied Statistics.*

Stone, R.A. (1988). Investigations of excess environmental risks around putative sources: Statistical problems and a proposed test. *Statistics in Medicine*, **7**, 649–60.

Wakefield, J.C. and Morris, S.E. (1999). Spatial dependence and errors-in-variables in environmental epidemiology. In: *Bayesian Statistics 6* (J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds.), pp. 657–684, Oxford University Press, Oxford.

Wakefield, J.C. and Morris, S.E. (1999). The Bayesian modelling of disease risk in relation to a point source. *Journal of the American Statistical Association*, **96**, 77–91.

Wakefield, J.C. and Salway, R. (2001). A Statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society, Series A*, **164**, 119–137.

White, J.E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, **115**, 119–128.

Wilkinson, P., Thakrar, B., Shaddick, G., Stevenson, S., Pattenden, S., Landon, M., Grundy, C. and Elliott, P. (1997). Cancer incidence around the Pan Britannica industries pesticide factory, Waltham Abbey. *Occupational and Environmental Medicine*, **54**, 101–107.

Wilkinson, P., Thakrar, B., Walls, P., Landon, M., Falconer, S., Grundy, C., and Elliott, P. (1999). Lymphohaematopoietic malignancy around all industrial complexes that include major oil refineries in Great Britain, *Occupational and Environmental Medicine*, **56**, 577–580.

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