

# Sensitivity Analyses for Ecological Regression

Jon Wakefield

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

## Summary

In many ecological regression 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 data since they arise from within-area variability in exposures/confounders. The potential for residual spatial dependence that is due to unmeasured confounders and/or data anomalies with spatial structure, may also be considered though often will be of secondary importance when compared to the likely effects of unmeasured confounding and within-area variability in exposures/confounders. Methods for addressing the sensitivity to these issues are described, along with an approach for assessing the implications of spatial dependence. An ecological study of the association between myocardial infarction and magnesium is critically re-evaluated to determine potential sources of bias. It is argued that the sophistication of the statistical analysis should not outweigh the quality of the data, and that finessing models for spatial dependence will often not be merited in the context of ecological regression.

*Keywords:* Confounding; Ecological fallacy; Spatial dependence; Spatial epidemiology; Within-area variability.

# 1 Introduction

The aim of this paper is to discuss a number of issues relating to ecological regression studies in the context of environmental epidemiology in which the association between risk and exposure is examined across areas. Such studies utilize data at the level of the group rather than the individual, have a long history in epidemiology (Morgenstern, 1998), and are appealing since they use routinely-available data, and can exploit large exposure contrasts. However, they are controversial due to a variety of difficulties that are summarized under the umbrella term of *ecological bias*. Incorrect conclusions may be reached in such studies for a variety of reasons 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* and *effect modification* (Greenland, 1992). The analysis of ecological data in the situation in which the groups are geographical areas is also complicated by the potential presence of spatial dependence in the residuals. There is a vast literature on ecological bias and the ecological fallacy in the epidemiological literature, e.g. Greenland and Robins (1994). It is the aim of this paper to review statistical models that are used for ecological data, and to suggest methods for assessing the sensitivity of inference to confounding and within-area variability in exposures. Unfortunately these issues have received little attention in the statistical literature where more effort has been placed on the consideration of covariance models for spatial dependence. We argue that the specific model used is of secondary importance when compared to the effects of confounding and within-area variability in exposure. It may also be misleading to attempt to account for spatial dependence if the exposure has spatial structure since some of the association may be absorbed into the spatial random effects. The sparsity of cases and the presence of errors in the disease and population counts also means that spatial modeling will often be driven by prior assumptions that are uncheckable.

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. For example, a number of point source studies have been carried out by the Small Area Health Statistics Unit (Elliott et al. 1992) in the UK and have reported excesses in risk at source in the range 0.1–1.0 that is, relative risks of 1.1–2.0. Wakefield and Morris (2001, p. 85), who reviewed the relevant literature in order to formulate an informative prior for the parameters of a distance/risk model, provide references to support this range. Such values are consistent with the value of 1.5 that is quoted by Pekkanen and Pearce (2001) as being typical of environmental studies. These relative risks must be viewed in light of the fact that, in particular for cancer and heart disease, 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 use individual-level risk factor data (though stratification by age, gender and socio-economic status is routinely carried out). 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 detrimental effect, but also on studies that reveal no association.

The structure of this paper is follows. In Section 2 we introduce a specific example to motivate the ideas of the paper. In Section 3 we describe notation and review a number of approaches

that are currently used to analyze ecological data, and in Section 4 we discuss the likely effects of spatial dependence. Section 5 considers how sensitivity to unmeasured confounding may be addressed in ecological studies. In Section 6 the sensitivity to aggregation of a nonlinear exposure/risk model in the presence of within-area variability in exposure is considered. In Section 7 we return to the magnesium example, and Section 8 provides a concluding discussion.

## 2 Magnesium Study

Here we re-analyze a subset of data analyzed substantively by Maheswaran et al. (1999) to examine the association between mortality from acute myocardial infarction (MI) and magnesium in drinking water in the years 1990–1992, in the north west of England. It has been hypothesized that although magnesium does not reduce the incidence of MI, it does protect against death from MI (Comstock, 1979). In the data we analyze, which are available in a region approximately 50 miles in the North-South direction and 44 miles in the East-West direction, exposure measures of magnesium are available from 225 water supply zones, which contain a maximum of 50,000 people, and provide the ecological level of the analysis. Specifically, magnesium concentrations were measured in the domestic water supply, with multiple measurements (median 5) per water zone (unfortunately the locations within water zones at which measurements were made is unknown). Measurements are also available for potential water constituent confounders calcium and fluoride. Expected counts for each water zone were obtained with standardization at the level of the enumeration district (which contain on average 400 people) for age, sex and the census-derived Carstairs measure of socio-economic status that combines the percentages in each area of individuals with no car, individuals in overcrowded households, with household head in social class IV or V, and unemployed men (Carstairs and Morris, 1991). The latter may be seen as acting as a

very rough surrogate for the aggregate lifestyle characteristics (e.g. diet, alcohol, smoking) of the individuals within the area. The expected and observed numbers across water zones have quartiles (80, 138, 198) and (82, 141, 205), respectively. Since the original water constituent measurements are highly skewed, we take the natural logarithm. For example Northings (which is defined in this study as the  $y$ -coordinate of the spatial coordinates of the water zone centroid), which has obvious spatial structure, is sometimes used as a surrogate for the effects of ultra-violet radiation from sunlight, or for temperature. The decision to include such a variable can often change the substantive conclusions of a study; this was the case in the magnesium study, see Table 2 of Maheswaran et al. (1999). The MLE of the relative risk is given by the observed/expected ratio, or Standardized Mortality Ratio (SMR), and in Figures 8(a)–(d) the log SMR is plotted versus the area-level means and shows no apparent ecological association with any of the water constituents or Northings. Figures (e)–(h) give pairwise plots of selected variables and show that, in particular for magnesium and calcium there is high correlation between the water constituents.

### 3 Statistical Framework

#### 3.1 Conventional Approaches

We consider a study area  $A$  that may be partitioned into sub-areas  $A_i$ ,  $i = 1, \dots, n$ , 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  $c$ ,  $c = 1, \dots, C$ . We let  $Y_i$  denote the observed number of cases, and  $X_i$ , an area-level measure of exposure, in area  $i$ ,  $i = 1, \dots, n$ . Known confounders are controlled for by calculating expected numbers  $E_i = \sum_c N_{ic} p_c$ , with  $p_c$  the “reference” probability of disease in stratum  $c$ . A basic model for a rare and non-infectious disease is to assume  $Y_i | R_i \sim \text{Poisson}(E_i \times R_i)$ , where  $R_i$  is the aggregate *relative*

*risk* associated with area  $i$ . Both imprecision due to small numbers, and overdispersion (that may be due to unmeasured risk factors with or without spatial dependence and/or data anomalies), may be addressed via the introduction of random effects. In the context of disease mapping Besag, York and Mollié (1991) proposed a model that included both spatially structured and unstructured random effects. An ecological regression form of this model was used by Clayton, Bernardinelli and Montomoli (1993) and is given by

$$\log R_i = \beta_0 + \beta_1 X_i + T_i + S_i, \quad (1)$$

where  $T_i | \tau^2 \sim N(0, \tau^2)$  denote unstructured (independent) random effects, and  $S_i$  random effects with spatial structure. Besag et al. (1991) modeled the latter using the intrinsic conditional autoregressive Markov random field specification with the conditional distribution  $S_i | S_j, j \in \partial i, \sigma^2 \sim N(\bar{S}_i, \sigma^2/m_i)$ , where  $\partial i$  represents the indices of a set of “neighboring” areas,  $m_i$  is the number of such neighbors, and  $\bar{S}_i$  is the mean of these neighbors.

A simple approach to ecological inference is possible for studies carried out over large geographical scales in which the number of counts in each area is not small. In this case the linear model  $\log \text{SMR}_i = \beta_0 + \beta_1 X_i + \epsilon_i$ , may be used, with spatial dependence allowed for through a covariance model for the collection  $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$ . Cook and Pocock (1982) assume the model  $\epsilon | \sigma^2, \rho \sim N_n\{\mathbf{0}_n, \Sigma(\sigma^2, \rho)\}$  where  $\mathbf{0}_n$  denotes an  $n \times 1$  vector of zeroes, and  $\Sigma(\sigma^2, \rho)$  is an  $n \times n$  variance-covariance matrix with  $(i, j)$ -th element

$$\Sigma_{ij} = \sigma^2 \exp(-d_{ij}\phi) = \sigma^2 \times \rho^{d_{ij}}, \quad (2)$$

where  $d_{ij}$  is the distance between the centroids of areas  $i$  and  $j$ , and  $\phi > 0$ , anticipating positive dependence.

## 3.2 Individual-Level Models

We now describe a hypothetical individual-level model where, for clarity, we assume there are no stratum. We let  $Y_{ij} = 0/1$  represent the event that individual  $j$  of area  $i$  is a non-case/case. We denote exposures of interest by  $X_{ij}$ , and confounders by  $U_{ij}$ . We then have the *individual-level model*  $E(Y_{ij}|X_{ij}, U_{ij}) = p(X_{ij}, U_{ij})$  and  $Y_{ij}|X_{ij}, U_{ij} \sim \text{Bernoulli}\{p(X_{ij}, U_{ij})\}$ . We may have an *additive model*

$$E(Y_{ij}|X_{ij}, U_{ij}) = p(X_{ij}, U_{ij}) = \beta_0 + X_{ij}\beta_1 + U_{ij}\beta_2, \quad (3)$$

in which  $\beta_1, \beta_2$  represent *risk differences*, or a *multiplicative model*

$$E(Y_{ij}|X_{ij}, U_{ij}) = p(X_{ij}, U_{ij}) = \exp(\beta_0 + X_{ij}\beta_1 + U_{ij}\beta_2), \quad (4)$$

where  $e^{\beta_1}, e^{\beta_2}$  are *relative risks*. Here and throughout we assume there are no contextual effects so that it is only an individual's risk factors that are relevant, and not the risk factors of other individuals in the area. We have also assumed that the risk summaries are constant across areas which is unlikely to be realistic, but in a rare-disease situation the lack of cases prevents the estimation of area-specific risk differences/relative risks.

## 4 Spatial Dependence

In the statistical literature, much attention has been given to the modeling of spatial dependence for aggregate data (e.g. Cressie, 1993; Best, Ickstadt and Wolpert, 2000 and references there-in). The motivation for this in an ecological regression context has been that ignoring positive spatial dependence (that is, positive dependence between the residuals of “neighboring” areas) will lead to *underestimation* of standard errors on regression coefficients, and will prevent “confounding by location” (Clayton, Bernardinelli and Montomoli, 1993), since often the exposure of interest will have spatial structure. In this section we examine the

effects of positive dependence in a simple time series context, and illustrate that the dependence need not lead to underestimation of the standard errors (though this will often be the case). Another important consideration is whether the inclusion of residuals with spatial structure will prevent confounding, or in fact distort the association.

We suppose for simplicity that data are collected across time at regular intervals, and that the model is given by

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t, \quad (5)$$

where the error terms  $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$  have distribution  $\epsilon \sim N_n(\mathbf{0}_n, \mathbf{W}\sigma^2)$  with

$$W_{st} = \rho^{|s-t|}, \quad s, t = 1, \dots, T; s \neq t. \quad (6)$$

Ignoring the temporal dependence and applying OLS gives  $\widehat{\text{var}}_O(\hat{\beta})$  proportional to  $(\mathbf{X}'\mathbf{X})^{-1}$ ; applying GLS gives  $\widehat{\text{var}}_G(\hat{\beta})$  proportional to  $(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}$ . In Table 1 we examine  $\widehat{\text{var}}(\hat{\beta}_1)$  with  $\mathbf{W}$  given by (6). We present the ratio of the standard errors, under a range of distributions for  $X_t$ . The first block of the table considers a variable that is independent of time, while blocks two and three consider  $X_t$  structured in time, with block two having  $X_t$  quickly varying, and block three  $X_t$  more slowly varying. We see that under-/over-estimation of standard errors depends critically on the distribution of the covariate across time relative to the size of the temporal dependence. As seen in block one, when  $X_t$  is unstructured, then the GLS standard error will be smaller than the naïve (OLS) estimate, contrasting with the commonly-held view. The intuitive reason for this is that observations close together are similar in all respects other than  $X_t$ , due to the positive dependence, and hence efficient estimation (that is smaller standard errors) will result. If  $X_t$  is more slowly varying then the standard errors will be underestimated by the naive analysis, in-line with the usual intuition. As  $\rho$  increases the amount of information decreases reflected in the standard errors under GLS. If the covariate and the residuals are changing on similar scales then there are competing explanations for the data which is reflected in increased standard errors. Heagerty and

Lumley (2000) considered both temporal and spatial situations, and observed this behavior with Gaussian and Poisson models. This phenomena has close connections with longitudinal studies in which the efficiency of a study depends on whether the covariate is changing within a person, and on the strength of the within-person correlation (Diggle, Liang and Zeger, 1994, Chapter 2).

In general the likely effect of spatial dependence may be investigated in an initial exploratory step. In an ecological setting, if a linear model is appropriate, then  $\mathbf{W}$  may be constructed from the geography, with the strength of dependence being governed by the distance between area-centroids. The matrix  $(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}$  may then be evaluated for plausible values of  $\rho$ , using the between-area means of the exposures to determine  $\mathbf{X}$ . The sensitivity of the variance of the estimator will aid in determining how much effort should be put into modeling the spatial dependence. This decision may also be addressed when interval estimates of  $\beta_1$  are obtained under a non-spatial model, to see if the form of spatial dependence is likely to alter the substantive conclusions. Many environmental exposures will produce area means that vary slowly over space, which suggests that standard errors will be underestimated when spatial dependence in the residuals is not included, though the extent of the underestimation may be small.

In terms of the utility of a particular ecological study, in addition to the conventional need for large exposure contrasts across the study region, we would add consideration of the *distribution* of exposure across the region, relative to the extent of spatial dependence. In particular, quickly-changing mean exposures across areas will have the joint benefits of producing reduced standard errors, and fewer difficulties due to confounding for the common situation in which confounders may have large-scale spatial pattern. Unfortunately a quickly-varying exposure is also likely to exhibit greater within-area variability and be more susceptible to pure-specification bias (which is discussed in Section 6).

We stress that if the aim of the study is not ecological regression but rather to estimate area-level estimates of risk (which may be thought of as a prediction problem), for health services allocation for example, then spatial dependence is likely to have a large impact on estimates and greater care may be merited with spatial modeling.

## 5 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). We stress that in this section we are interested in bias, in the presence of unmeasured variables, confounders or otherwise. The distribution of the data is in general no longer Poisson, and in particular the variance is larger than the mean. In practice, one indicator of the presence of unmeasured variables is the extent of overdispersion. Simple methods for adjusting the variance of the estimating include quasi-likelihood and sandwich estimation though neither can account for spatial dependence unless there is replication across time. Alternatively, random effects models, such as those described in Section 3.1 may be fitted.

### 5.1 Additive Model

We first consider model (3). In the following let  $X_i = E(X_{ij}|i)$  and  $U_i = E(U_{ij}|i)$ , denote the area level means of confounder and exposure respectively. Throughout this section we use iterated expectation to obtain the ecological model when only area-level data are available. If there is no between-area confounding (so that  $X_i$  and  $U_i$  are independent) then

$$E(Y_{ij}|X_i) = \beta_0^* + \beta_1 X_i, \tag{7}$$

where  $\beta_0^* = \beta_0 + \beta_2 U_i$ , and so, even if there is within-area confounding, no bias will result. It also follows that regardless of the within-area behavior, 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 in some situations.

### 5.1.1 Binary variables

Suppose  $X$  and  $U$  are both binary so that  $X_i = \Pr(X_{ij} = 1|i)$  represents the proportion exposed and similarly  $U_i = \Pr(U_{ij} = 1|i)$ . We define  $\Pr(U_{ij} = 1|X_{ij} = x, i) = P_{ix}$ ,  $x = 0, 1$ ,  $i = 1, \dots, n$ , so that  $U_i = \sum_{x=0}^1 \Pr(U_{ij} = 1|X_{ij} = x, i) \Pr(X_{ij} = x|i) = P_{i0} + (P_{i1} - P_{i0})X_i$ . Under this model no confounding corresponds to  $P_{i0} = P_{i1} = P_i$  and if  $P_{i0} \neq P_{i1}$  we have confounding both within and between areas. We have

$$E(Y_{ij}|X_i) = \beta_0 + \beta_2 P_{i0} + \{\beta_1 + \beta_2(P_{i1} - P_{i0})\}X_i = \beta_{i0}^* + \beta_{i1}^* X_i, \quad (8)$$

where  $\beta_{i0}^* = \beta_0 + \beta_2 P_{i0}$  and  $\beta_{i1}^* = \beta_1 + \beta_2(P_{i1} - P_{i0})$ , showing that both an area-specific intercept, and effect modification by area have been induced. If  $\Delta P = P_{i1} - P_{i0}$  is constant across areas then there is no effect modification by area. This development provides a justification for the use of random effects models in an ecological context.

An obvious way to examine sensitivity to between-area confounding is to assume a constant  $\Delta P$ , so that the relationship between confounder and exposure does not change with  $i$ , though the baseline prevalence of the confounder may vary, and we obtain  $\beta_1^* = \beta_1 + \beta_2 \Delta P$ . 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, in which case  $P_{ix} = \Pr(\text{deprived}|\text{radon } x, i)$ ,  $x = 0, 1$ . Suppose  $E[P_{i0}] = 0.1$  and  $\Delta P = 0.2$  so that the probability of being deprived is on average three times higher if resident in a high radon location. If there is truly no effect of radon on lung cancer ( $\beta_1 = 0$ )

and  $\beta_2 = 0.1$  (so that deprived individuals have a risk 0.1 greater than affluent individuals), but we do not measure deprivation, then we would incorrectly estimate the risk difference as  $\beta_1^* = 0.02$ . Conversely if there is no observed effect of  $\beta_1^* = 0$ , this could be hiding a true effect of  $\beta_1 = 0.02$  if  $\beta_2 = 0.05$ ,  $E[P_{i0}] = 0.5$ ,  $\Delta P = -0.4$ .

We now briefly consider the situation in which we have an interaction, i.e.  $E(Y_{ij}|X_{ij}, U_{ij}) = \beta_0 + \beta_1 X_{ij} + \beta_2 U_{ij} + \beta_3 X_{ij} U_{ij}$ . Suppose we observe area proportions,  $X_i, U_i$ , then  $E(X_{ij} U_{ij} | X_i, U_i) = P_{i1} X_i$ , and model (8) leads to  $E(Y_{ij} | X_i, U_i) = \beta_0 + (\beta_1 + \beta_3 P_{i1}) X_i + \beta_2 U_i$ , which allows sensitivity to be addressed. Laserre et al. (1999) examine this case and advocate the approximation  $E(X_{ij} U_{ij} | X_i, U_i) \approx X_i U_i$  which corresponds to independence, that is,  $P_{i1} = U_i$  and makes an attempt to acknowledge the presence of  $U$  (which they show, in a variety of simulations, to be better than ignoring this term).

### 5.1.2 Normal variables

Now consider the case of no interaction and a continuous exposure and 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

$$\begin{bmatrix} X_i \\ U_i \end{bmatrix} \sim N \left( \begin{bmatrix} \mu^x \\ \mu^u \end{bmatrix}, \begin{bmatrix} \Sigma^x & \Sigma^{xu} \\ \Sigma^{ux} & \Sigma^u \end{bmatrix} \right), \quad (9)$$

where  $\Sigma^{ux} = \rho(\Sigma^x \Sigma^u)^{1/2}$ . If we regress on  $X_i$  only via (7) we obtain  $\beta_0^* = \beta_0 + \beta_2 \{ \mu^u - \mu^x (\Sigma^u / \Sigma^x)^{1/2} \rho \}$  and

$$\beta_1^* = \beta_1 + \beta_2 (\Sigma^u / \Sigma^x)^{1/2} \rho. \quad (10)$$

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 with

greater bias if the confounder has large variance relative to the exposure (providing further justification of the requirement for large exposure contrasts across areas). This model may easily be extended to multiple confounders, as described in the next section.

## 5.2 Multiplicative Model

Now suppose we have the model (4). In the absence of between-area confounding, within-area confounding will lead to bias, in contrast to the additive model case.

### 5.2.1 Binary variables

We first suppose that  $X_{ij}$  and  $U_{ij}$  are binary and again assume  $P_{ix} = \Pr(U = 1 | X_{ij} = x, i)$ .

We will examine the bias that results when the model

$$E(Y_{ij}|X_i) = \exp(\beta_0^* + \beta_1^* X_i), \quad (11)$$

is fitted, which is equivalent to  $E(Y_{ij}|X_i) = \alpha_0^* + \alpha_1^* X_i$ , where  $\alpha_0^* = \exp(\beta_0^*)$  and  $\alpha_1^* = \exp(\beta_0^*)\{\exp(\beta_1^*) - 1\}$  and the relative risk  $\exp(\beta_1^*) = 1 + \alpha_1^*/\alpha_0^*$ .

We obtain  $E(Y_{ij}|X_i) = \exp(\beta_{i0}^* + \beta_{i1}^* X_i)$ , where  $\beta_{i0}^* = \beta_0 + \log[(1 - P_{i0}) + P_{i0} \exp(\beta_2)]$ , and  $\beta_{i1}^* = \beta_1 + \log\{(1 - P_{i1}) + P_{i1} \exp(\beta_2)\} - \log\{(1 - P_{i0}) + P_{i0} \exp(\beta_2)\}$ . Lin, Psaty and Kronmal (1999) obtain the equivalent of this form when  $X_i = 0/1$  is constant within areas. An observed relative risk of  $e^{\beta_1^*} = 1.19$  could be obtained with  $\beta_1 = 0$ ,  $e^{\beta_2} = 1.5$ ,  $P_0 = 0.1$  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).

### 5.2.2 Normal variables

Now consider continuous exposure/confounders and suppose that (9) applies. If we assume that there is no within-area variability in exposures or confounders then the coefficients of model (11) are again given by (10).

The extension to multiple confounders is straightforward. Suppose  $E(Y_i|X_i, U_{1i}, \dots, U_{Ci}) = \exp(\beta_0 + \beta_1 X_i + \beta_2 \sum_{c=1}^C U_{ci})$ . Then under multivariate normality of  $X_i, U_{1i}, \dots, U_{Ci}$  with  $\text{corr}(X_i, U_{ci}) = \rho$  and  $\text{var}(X) = \text{var}(U_{ci})$  we have  $E(Y_i|X_i) = \beta_0^* + \beta_1^* X_i$ , where  $\beta_1^* = \beta_1 + \beta_2 C \rho$ , so that if  $\beta_1 = 0$  we have  $\beta_1^* = \beta_2 \times C \times \rho$  showing the exact interplay between number of confounders, strengths of dependence and confounder association, and resultant estimate (when there is no association).

Now consider the situation in which we have within and between area confounding but we measure both  $X_i$  and  $U_i$ . If we assume

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

then we obtain

$$E(Y_{ij}|X_i, U_i) = \exp\{\beta_1 X_i + \beta_2 U_i + (\beta_1^2 \Sigma_i^x + \beta_2^2 \Sigma_i^u + 2\beta_1 \beta_2 \Sigma_i^{xu})/2\}. \quad (12)$$

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 independent of  $X_i$  then these terms will be absorbed into the intercept and no bias will result.

For both additive and multiplicative models the developments of this section show that unmeasured variables naturally leads to a random effects formulation with the distribution of the covariate across areas determining the distribution of the random effects. The inclusion of random effects in a model cannot in general control for confounding, however.

## 6 Pure Specification Bias

In this section we assume for simplicity that there are no confounders and consider in isolation *pure specification bias* which is the effect of aggregation of the individual exposure/risk model.

We assume a univariate continuous exposure and that within area  $i$ ,  $X|\phi_i \sim_{i.i.d.} f(\cdot|\phi_i)$  where  $\phi_i$  denotes a set of parameters that characterize the exposure. Then, for individual  $j$  in area  $i$ ,  $j = 1, \dots, N_i$ ,  $Y_{ij}|\beta, \phi_i \sim_{i.i.d.} \text{Bernoulli}\{p(\phi_i)\}$ , where  $p(\phi_i) = E_{X|\phi_i}\{p(X)\} = \int p(x)f(x|\phi_i)dx$ . If exposures are independent within areas, and the outcome is rare, then  $Y_i|\beta, \phi_i \sim \text{Poisson}\{n_i p(\phi_i)\}$ , 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  $p(\phi_i) = \beta_0 + \beta_1 X_i$ . For the multiplicative model  $p(x) = \exp(\beta_0 + \beta_1 x)$  we have

$$p(\phi_i) = \exp(\beta_0)E\{\exp(X\beta_1)\} \quad (13)$$

(Richardson, Stucker and Hemon 1987) where the expectation is with respect to  $X|\phi_i$ . For the case  $X|\phi_i \sim N(X_i, \Sigma_i^x)$  with  $\phi_i = (X_i, \Sigma_i^x)$  where  $X_i = E(X|i)$  and  $\Sigma_i^x = \text{var}(X|i)$  we obtain  $p(\phi_i) = \exp(\beta_0 + \beta_1 X_i + \beta_1^2 \Sigma_i^x / 2)$ . The key to understanding bias here is to view the variance as an additional variable with positive effect (since  $\beta_1^2 / 2 > 0$ ); no bias results if the means and variances are unrelated and, in particular, if  $\Sigma_i^x$  are constant across areas. We describe a simple method for determining the extent of the bias using ideas from the last section. Suppose that across areas we have a linear relationship between the variance and the mean,  $\Sigma_i^x = a + bX_i$ , then model (11) holds with  $\beta_0^* = \beta_0 + a\beta_1^2/2$ ,  $\beta_1^* = \beta_1 + b\beta_1^2/2$ . Hence if  $\beta_1 > 0$  and, as we might expect, the variance increases with the mean ( $b > 0$ ), then ignoring within-area variability will lead to overestimation of  $\beta_1$ ; if  $\beta_1 < 0$  so that the exposure is protective, the estimate of the effect may even change sign. It has been recognized that many ecological studies find larger effects than their individual-level counterparts (see Maheswaran et al. 1999 for references to support this in the context of magnesium/MI). To

assess sensitivity, for an observed  $\widehat{\beta}_1^*$  and plausible  $b$  (perhaps based on observed means and variances of the exposure across areas), the quadratic  $b\beta_1^2/2 + \beta_1 - \widehat{\beta}_1^* = 0$ , may be solved for  $\beta_1$ .

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 for each area a measure of the spread in exposure is available. The choice  $X|\phi_i \sim U(X_i - c_i, X_i + c_i)$ , with  $\phi_i = (X_i, c_i)$ , gives  $p(\phi_i) = \exp(\beta_0 + \beta_1 X_i)(e^{\beta_1 c_i} - e^{-\beta_1 c_i})/2c_i\beta_1$ , for  $\beta_1 \neq 0$ . Hence there is no bias if the spread  $c_i$  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  $\phi_i$ . Wakefield and Salway (2001) show that for small within-area samples the estimation of the variance in particular is highly unstable and can lead to inaccurate inference. To alleviate the instability, one possibility is to model the variance as a smooth function of the mean. Prentice and Sheppard (1995) provide an alternative method in which (13) is estimated nonparametrically using within-area samples of exposures, with (non-spatial) random effects. Guthrie, Sheppard and Wakefield (2002) extend this approach to include spatial random effects within a hierarchical model.

This consideration of pure specification bias indicates that random effects could also be representing within-area variability in exposures/confounders; for example  $T_i$  and  $S_i$  in (1) may be accommodating 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 (12). 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 5 we

described models that were marginalized across unmeasured area-level variables, whereas in the random effects formulation 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.

## 7 Magnesium Study

In this section we return to the ecological regression study introduced in Section 2. Table 2 reports summaries from a variety of models that were fitted to these data. Magnesium is the variable of primary interest, we also examine Northings as it provides an illustration of a variable with strong spatial structure. We first note that in none of the models for either magnesium or Northings is there evidence of an association but the extent of overdispersion (variance almost three times the mean) indicates the existence of important unmeasured variables (or data anomalies), and hence the possibility for unmeasured confounding which could be masking a true association.

### *Spatial Dependence*

We first calculate the ratio of standard errors  $\{(\mathbf{X}'\mathbf{X})_{22}^{-1}/(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})_{22}^{-1}\}^{1/2}$  where  $\mathbf{X}$  is the  $225 \times 2$  design matrix containing a column of 1's, and the vector of mean log magnesium measurements by area, and  $\mathbf{W}$  contains  $(i, j)$ -th element  $\rho^{d_{ij}}$  where  $d_{ij}$  is the distance between the centroids of water zones  $i$  and  $j$ . With  $\rho = 0.1$  which corresponds to the residuals having correlation 0.5 at points at a distance of 8 miles apart (recall the study region is approximately  $50 \times 44$  miles), we obtain a ratio of 1.4 which indicates that adjusting for spatial dependence will not make much difference for magnesium. Figure 1(h) shows that there is only a slight spatial trend with higher levels of magnesium in the North. By contrast, for Northings the ratio is 0.3, indicating that standard errors would be three times larger.

Table 2 illustrates that inclusion of spatial random effects, does indeed have a considerably greater impact on the precision of the estimate for Northings. For the multivariate spatial specification convergence difficulties were encountered, particularly for the analysis including Northings (for which  $N(0, 1)$  priors were used for  $\alpha$  and  $\beta$ ). This model was also hugely more computationally intensive than the intrinsic CAR model.

### *Confounding*

Since a multiplicative (loglinear) model was used in the original analysis, we first use the results of Section 5.2 to assess the sensitivity to unmeasured confounding. Figure 1(i) shows that the assumption of between-area normality of log magnesium is a reasonable approximation here. The observed relative risk is equal to  $\exp(\hat{\beta}_1^*) = e^{0.0075} = 1.008$  but (from (10)) if a single confounder with the same between-area variability were unmeasured then the true reduction in risk would be 4% if the unmeasured confounder produced an increase in risk of 10%, and had correlation 0.5 with log magnesium. Four confounders, each with correlation of 0.2 with log magnesium, and each producing a 4% increase in relative risk would hide a 2% reduction due to (log) magnesium. In an ecological study such as this there are many potential confounders including unmeasured water constituents, or inadequate control for lifestyle characteristics such as diet, alcohol and smoking by the socio-economic status index. The inclusion of calcium and fluoride gave no change in the conclusions with respect to the magnesium/MI association, and explained virtually none of the unmeasured variability.

### *Within-Area Variability*

The interquartile range of the sample within-area variances is (0.53,0.69) while the range of mean (log) concentrations is approximately -2 to 2, and in the original study the within-area variability of log magnesium was found to be approximately normal. The within-area variability in log magnesium accounted for 24% of the total variability. Assuming  $\Sigma_i^x \approx a + bX_i$  leads to plausible values of  $b$  in the range 0.05 to 0.10. Under this model and

with  $\hat{\beta}_1^* = 0.0075$  and, for example  $b = 0.08$ , the relative risk estimate increases to 1.078 under the more plausible of the two roots). Such a large increase could be moved in the opposite direction if confounding has caused the observed association to be of the wrong sign.

### *Measurement Error*

Measurement error is unlikely to cause significant bias here since the mean area exposure measures are based on multiple concentration measurements per area.

For this study, we conclude that unmeasured confounding and within-variability in exposure are the likeliest sources of bias, and could be substantial.

## **8 Discussion**

In this paper we have considered a variety of issues relating to the analysis of ecological regression data. A key point is that while acknowledging spatial dependence in residuals may be beneficial in order to obtain an appropriate standard error on regression coefficients, this enterprise will often be of secondary importance when compared to biases that are due to data anomalies, unmeasured confounding and within-area variability in exposures and confounders. Methods to address the likely extent of these biases have been described, along with an approach for determining the possible effect of spatial dependence. In many cases refining the model for spatial dependence will not be merited. If the estimated association changes when spatial residuals are introduced then care in interpretation is required as the estimate is based on the local association between risk and exposure which may not be appropriate.

The considerations of this paper indicate that in a well-designed ecological study in which known confounders are collected and a multiplicative model is used, 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 available.

We finally emphasize that the comments of this paper are specifically addressed to ecological regression studies, for disease mapping (which is more a problem of prediction), and studies in which point data are obtained (for example in a case-control design), then the benefits of spatial modeling may be more substantial.

## Acknowledgements

The author would like to thank Patrick Heagerty for useful discussions concerning the material in Section 4, and the editor, associate editor, and a referee for constructive comments. The author thanks the Office for National Statistics for the data on acute myocardial infarction, and the Estimating with Confidence Project for the population data.

## References

- 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., Ickstadt, K. and Wolpert, R.L. (2000 ). Spatial Poisson regression for health and exposure data measured at disparate spatial scales. *Journal of the American Statistical Association*, **95**, 1076–1088.

- Carstairs, V. and Morris, R. (1991). *Deprivation and Health in Scotland*, Aberdeen: Aberdeen University Press.
- Clayton, D.G., Bernardinelli, L. and Montomoli, C. (1993). Spatial correlation in ecological analysis. *International Journal of Epidemiology*, **22**, 1193–1202.
- Comstock, G.W. (1979). Water hardness and cardiovascular disease. *American Journal of Epidemiology*, **110**, 375–400.
- Cook, D.G. and Pocock, S.J. (1983). Multiple regression in geographical mortality studies, with allowance for spatially correlated errors. *Biometrics*, **39**, 361–371.
- 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.
- Cressie, N.A.C. (1993). *Statistics for Spatial Data, Revised Edition*, John Wiley and Sons.
- Diggle, P.J., Liang, K.Y., and Zeger, S.L. (1994). *Analysis of Longitudinal Data*, Oxford Science Publications.
- Elliott, P., Westlake, A., Hills, M., Kleinschmidt, I., Rodrigues, L., McGale, P., Marshall, K., and Rose, G. (1992). 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.
- Greenland, S. (1992). Divergent biases in ecologic and individual-level studies, *Statistics in Medicine*, **11**, 1209–23.
- Greenland, S. and Robins, J. (1994). Ecological studies-biases, misconceptions and counterexamples. *American Journal Epidemiology*, **139**, 747–760.
- Guthrie, K.A., Sheppard, L. and Wakefield, J.C. (2002). A hierarchical aggregate data model

with spatially correlated disease rates. To appear in *Biometrics*.

Heagerty, P. and Lumley, T. (2000). Window subsampling of estimating functions with application to regression models. *Journal of the American Statistical Association*, **95**, 197–211.

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.

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.

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

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

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.

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 sta-

tus in estimates of the associations between occupation and cancer. *Journal of Occupational Medicine*, **30**, 617–625.

Wakefield, J.C. and Morris, S.E. (2001). The Bayesian modeling 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.

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

**Figure 1:** Selected exploratory plots for the magnesium/heart disease example. Ecological associations between outcome and exposures at the water zone level: log SMR versus log magnesium (a), log calcium (b), log fluoride (c), Northings (d). Ecological relationships between variables at the water zone level: log magnesium versus log calcium (e), log magnesium versus log fluoride (f), log calcium versus log fluoride (g), log magnesium versus Northings (h). Normal scores plot for average log magnesium levels (i).

Distribution of exposure	$\rho$	$\tilde{\sigma}^2$	$\widehat{\text{var}}_O(\hat{\beta}_1)$	$\hat{\sigma}^2$	$\widehat{\text{var}}_G(\hat{\beta}_1)$	Ratio of se's OLS/GLS
$X_t \sim_{iid} N(0, 1)$	0.0	1.000	0.0369	1.000	0.0369	1.000
	0.1	1.000	0.0367	1.010	0.0362	1.007
	0.5	0.946	0.0351	0.993	0.0222	1.257
	0.9	0.589	0.0220	1.010	0.00412	2.311
$X_t = \{0, 0, 0, \dots, 1, 1, 1\}$	0.0	1.003	0.134	1.003	0.134	1.000
	0.1	0.986	0.131	1.000	0.159	0.908
	0.5	0.878	0.117	1.000	0.300	0.624
	0.9	0.397	0.0529	1.007	0.188	0.530
$X_t = \{1, 2, 3, \dots, n\}$	0.0	1.000	0.000445	1.000	0.000445	1.000
	0.1	0.983	0.000438	0.998	0.000531	0.908
	0.5	0.889	0.000396	1.000	0.00110	0.600
	0.9	0.369	0.000164	0.988	0.00217	0.275

Table 1: Reported standard error as a function of the distribution of  $X_t$ ,  $t = 1, 2, \dots, n = 30$ , and the strength of dependence  $\rho$ . The entries are averages over 1000 simulations. The variance of  $\hat{\beta}_1$  that would be reported under OLS is denoted  $\widehat{\text{var}}_O(\hat{\beta}_1)$ , while  $\widehat{\text{var}}_G(\hat{\beta}_1)$  represents the appropriate variance under the serial correlation model and GLS. The estimates of  $\sigma^2$  under OLS and GLS are denoted  $\tilde{\sigma}^2$  and  $\hat{\sigma}^2$ , respectively. The second block in the table corresponds to the case in which the first  $n/2$  time periods have exposure level zero, and the second  $n/2$  have level one. The ratio denotes the simple ratio of average standard errors, a ratio larger (smaller) than 1 indicates that OLS overestimates (underestimates) standard errors and so is conservative (anti-conservative).

EXPOSURE	MODEL	$\hat{\beta}_1$	Standard Error	Width of 90% interval	Non-spatial $\hat{\sigma}$	Spatial $\hat{\sigma}$
Magnesium	Poisson	0.0075	0.0069	0.0227	–	–
	Quasi-Likelihood	0.0075	0.0113	0.0371	$\hat{\kappa} = 2.72$	–
	Non-spatial r.e.	0.0066	0.0118	0.0387	0.1100	–
	Non-spatial+ICAR r.e.	0.0033	0.0147	0.0482	0.0225	0.1542
	Non-spatial+MVN r.e.	-0.0108	0.0160	0.0525	0.0227	0.1284
Northings	Poisson	-0.0079	0.0169	0.0555	–	–
	Quasi-Likelihood	-0.0079	0.0278	0.0915	$\hat{\kappa} = 2.72$	–
	Non-spatial r.e.	0.0061	0.0288	0.0943	0.1101	–
	Non-spatial+ICAR r.e.	0.0463	0.0867	0.2849	0.0222	0.1167
	Non-spatial+MVN r.e.	0.0433	0.1943	0.6022	0.0546	0.1769

Table 2: Inferential summaries for the acute myocardial infarction data under various models.

In the Poisson model, the variance is equal to the mean, in the quasi-likelihood approach  $\text{var}(Y) = \kappa \times E[Y]$  so that the variance is linear in the mean. In the Bayesian random effect (r.e.) models the variance is a quadratic function of the mean and so is not directly comparable with the quasi-likelihood approach. For the intrinsic CAR (ICAR) random effects model the empirical standard deviation of the random effects is reported. In the multivariate (MVN) spatial specification, for magnesium, the posterior median for  $\phi$  was 3.4 which corresponds to the spatial correlation dropping to 0.5 at a distance of 5.4 miles. A similar strength was found for the Northings analysis. The priors on the variances were taken to be inverse gamma with parameters 0.5, 0.005 (see Wakefield and Morris, 2001 for a discussion of this specification), and the prior for  $\phi$  was uniform on the range 0.35,308. The choice 0.35 allows a maximum correlation of 0.5 at 50 miles and 308 allows a minimum correlation of 0.01 at 0.4 miles, which corresponds to the shortest distance between centroids in the study.

