

# Bayesian hierarchical models in ecological studies

Sylvia Richardson<sup>1</sup>, Chris Jackson<sup>1</sup> and Nicky Best<sup>1</sup>

<sup>1</sup> Department of Epidemiology and Public Health, Imperial College, Norfolk Place, London, W2 1PG, U.K.

**Abstract:** The aim of this article is to set out a generic methodological framework for combining aggregated and individual level data for ecological inference and to illustrate some of the benefits and difficulties by simulations and on a case study.

**Keywords:** Bayesian models, hierarchical models, ecological inference, ecological fallacy, data integration

## 1 Introduction

Ecological regression studies have a long history in epidemiology (Doll, 1980) as they aim to exploit observed spatial variations in disease rates and investigate how these may be associated with area level measures summarising major determinants of health (see Elliott et al., 2000, for a wealth of examples). Similarly, in the social or political sciences, ecological inference has been used to tease out factors influencing for example voting behaviour or crime rates (King, 1997). Ecological studies typically use observational data collected on groups defined by geographical areas, such data routinely arising through various types of census, national registers or public health systems. In this article, we will illustrate the methods developed in the epidemiological context and discuss in particular a case study that is focussed on analysing the variation of limiting long-term illness (LLTI), a health outcome that is systematically recorded in the UK census and which gives a useful ‘health snap shot’ of small areas.

In epidemiology, there have been two main contexts where ecological studies have been used: environmental epidemiology where ‘physical’ risk factors such as air pollution, chemical contamination of water or soil, and background radiations are of interest (e.g. Best et al., 2001), and social health where the multiple aspects of deprivation and their influence on disease risk and mortality have been investigated (e.g. Ben-Shlomo et al., 1996). In both cases, it has been argued that characterising some risk factors at an area level can be beneficial. Indeed, using area data for measuring some environmental variables alleviates individual-level measurement error problems while providing a greater range of contrasts (Richardson and Monfort,

2000); in other cases, the factors of interest like deprivation are themselves only defined at a group level.

By their simple design, their use of routinely collected data and their exploitation of natural geographical contrasts, ecological regression studies are appealing. Nevertheless, their interpretation is far from straightforward due to the loss of information in only having aggregated data and no direct link between exposure and outcome. In particular, if the interest is in using such studies to make individual level inference on the effect of specific risk factors, then a number of issues arise that have been commonly referred to as *the ecological fallacy*. Under this general heading are grouped distinct and interlinked reasons for having a different exposure-outcome relationship between group level data and individual level data. The types of bias arising in ecological inference can be broadly categorised as: specification bias, effect modification and different aspects of confounding (Richardson et al. 1987, Piatadosi et al. 1988, Greenland and Morgenstern 1989, Richardson 1992, Greenland and Robins 1994, Wakefield 2003). Specification bias arises when there is a non linear relationship between outcome and exposure that is not properly accounted for in the specification of the form of the aggregated model. Confounders in ecological studies may be unmeasured area-level variables or factors which vary between individuals. These may confound the baseline disease risk for groups or the effect of the risk factor under study, creating interactions and effect modifications.

At the other end of the spectrum, observational data sets with individual level information are collected in various types of surveys (including samples of anonymized census records) and large cohort studies. These data sets provide valuable direct information on exposure-outcome relationships and on a set of confounders, but as they commonly only sample a small number of individuals in each group/geographical area, they typically have low power, in particular to investigate potential links with area level determinants and contextual effects. On the other hand, as discussed by Greenland (2001) and Sheppard (2003), from the analysis of area level data only, it is particularly difficult to distinguish between individual level and contextual risks. Hence, there is a strong argument for developing methods that aim to improve both ecological inference and small area inference from surveys by *combining* aggregate and individual-level survey data.

The aim of this article is to set out a generic methodological framework for doing this type of *data integration*, to illustrate some of the benefits in terms of control of ecological bias and variance reduction using a simple simulation set-up and finally to discuss some of the difficulties of performing in practice such data integration on a case study concerned with LLTI.

## 2 Study designs

Let  $j = 1, \dots, N_i$  denote individuals within groups  $i = 1, \dots, I$ , where  $i$  may index, for example, time units, geographical (spatial) units, socioeconomic

groups, etc. We denote the total population at risk, the number of cases in group  $i$  and group level exposure(s) by  $N_i$ ,  $y_i$  and  $Z_i$  respectively; individual dichotomous health outcomes and exposure(s) of interest by  $y_{ij}$  and  $X_{ij}$  respectively, and the baseline risk (possibly adjusted for known risk factors such as age and sex) by  $\lambda_{0ij}$  (individual-level) or  $\Lambda_{0i}$  (group level). Depending on the type and resolution of the available data, different designs have been distinguished when trying to study relationships between risk factors and health outcomes (Richardson and Best, 2003).

In an *individual design* both exposure and health outcome data are measured at the individual level on the same set of subjects and the interest is in estimating the individual effects linking  $X_{ij}$  and  $y_{ij}$ . In a pure *ecological design*, the inference is focussed on the area level relationship between  $Z_i$  and  $y_i$ . In a *mixed design*, group level outcomes are supplemented by some individual outcome data and different types of exposure data are also available. Mixed design analysis allows a richer range of questions to be explored, for example, the aims may be:

1. Assessing how much information there is in the aggregated data about the individual effects
2. Investigating how the inclusion of a small sample of individual data can improve ecological inference, and conversely how inclusion of aggregate data can reduce the mean square error from an analysis of individual level data alone
3. Investigating jointly the influence of individual and contextual effects on the health outcome.

### 3 Aggregation of individual level models

In general, the functional form of the local dependence between the probability  $p_{ij}$  of binary outcome for individual  $j$  in group  $i$  and an exposure  $X_{ij}$  can be represented as:  $p_{ij} = g(\lambda_{0ij}, X_{ij}, \beta)$ ,  $j = 1, \dots, N_i$ , where  $g(\cdot)$  is a suitable link function.

Assume that individual-level exposures and baseline risk are unknown and that their joint probability distribution in the group  $i$  is given by  $f_i(\lambda_0, X)$ . The marginal probability  $p_i$  of health outcome for any individual in the group  $i$  is then given by the integral of the individual's conditional outcome probability  $p_{ij}$  with respect to the joint within-group distribution of  $\lambda_0$  and  $X$ . Throughout, we shall make the additional simplifying assumption that the baseline risks and the exposures  $X$  are independent within the group, so that their joint distribution is simply the product  $f_i(\lambda_0)f_i(X)$  where  $f_i(\lambda_0)$  and  $f_i(X)$  denote respectively the within-group distributions of  $\lambda_0$  and  $X$ . We thus obtain the expression:

$$p_i = \iint g(\lambda_0, X, \beta) f_i(\lambda_0) f_i(x) d\lambda_0 dx \quad (1)$$

Then, we model the number of cases  $y_i$  in the group by

$$y_i \sim \text{Bin}(N_i, p_i). \quad (2)$$

Note that we are not insisting on the health outcome being rare and so a Poisson approximation is not warranted. Throughout we use the binomial model (2) for the likelihood of the aggregated data. A different likelihood that involves instead a convolution model that conditions on known individual level exposures has been discussed at length by Wakefield (2004). Several assumptions can be made in order to obtain an explicit expression for (1) that can be combined with the binomial model (2) in order to estimate the parameters of the link function  $g$ . We start by recalling some simple cases where explicit integration can be carried out in (1).

### 3.1 Linear dose-effect relationship

Suppose that

$$g(\lambda_{0ij}, X_{ij}, \beta) = \lambda_{0ij} + \beta X_{ij}, \quad (3)$$

This formulation supposes that suitable constraints are operating on the range of  $X$  and  $\beta$ , so that the function in (3) still defines a probability. Because of this, the linear model (3) has not been much used in epidemiology and we include it here for reference. From (1) and (3), we obtain:

$$p_i = E_i(\lambda_0) + \beta E_i(X) \quad (4)$$

where  $E_i(\lambda_0)$  and  $E_i(X)$  represent the mean baseline risk and mean exposure respectively in group  $i$ . Note that it is then easy to combine (2) and (4) to obtain direct estimates of  $\beta$  from aggregate data on the number of cases and mean exposure per area for each of the covariates. Note further that the linear form of (4) is preserved if an additional group level covariate effect: ‘ $+\gamma Z_i$ ’ is included in (3), with corresponding change to (4).

### 3.2 Exponential dose-effect relationship with normally distributed exposure

A common form of dose-effect relationship used in epidemiology assumes multiplicative effects of risk factors (here, for simplicity,  $X_{ij}$  is univariate):

$$g(\lambda_{0ij}, X_{ij}, \beta) = \lambda_{0ij} \exp(\beta X_{ij}). \quad (5)$$

Substituting in (1), we obtain:

$$p_i = E_i(\lambda_0) E_i(\exp(\beta X)), \quad (6)$$

an expression that does not involve directly  $E_i(X)$ . If  $X \sim N(m_i, s_i^2)$ , we have:  $E_i(\exp(\beta X)) = \exp(\beta m_i + 0.5\beta^2 s_i^2)$ . (Note that in the case of

multivariate exposures, the full variance-covariance matrix of the joint exposure distribution enters here). In general, we can evaluate the expression  $E_i(\exp(\beta X))$  for any distribution of  $X$  for which the moment generating function (Laplace transform) is known explicitly, as was noted in Richardson et al (1987). Substituting in (6) with  $m_i = E_i(X)$ , we obtain as aggregated form of (1) in the Gaussian case

$$p_i = E_i(\lambda_0) \exp(0.5\beta^2 s_i^2) \exp(\beta E_i(X)). \quad (7)$$

We see that in this case, the aggregated form for  $p_i$  is also a function of the within-area variance of the exposure variable (or the within-area covariance matrix for multiple exposures). Neglecting this term thus leads to a biased functional relation at the group level. This bias is often referred to as *specification bias* (see Richardson and Monfort, 2000, Wakefield, 2003). There are some situations where this bias can become negligible (Plummer and Clayton, 1996). This is the case if either (i)  $s_i^2$  is small, or (ii)  $s_i^2$  hardly varies with  $i$  and thus can be absorbed in a constant term. Essentially for (i) to hold, the exposure has to be nearly uniform over the group which is rarely the case, except if the group is small, whilst it is difficult for (ii) to hold if the mean  $m_i$  also varies between groups. Thus, it is important to have some knowledge of the within-area (co)variance of the exposure(s), and to input this into the specification of the dose-effect relationship at the group level, especially since the terms  $s_i^2$  and  $E_i(X)$  are often correlated. Note that calculations in the Gaussian case can be extended to other within-area distributions for  $X$  (Wakefield and Salway, 2001). However, in general, the moment generating function of  $X$  will involve higher order moments of the within-area distribution of  $X$  and these would be hard to estimate. When partial information on moments or on bounds of the covariate distribution is available, some progress can also be made by resorting to maximum entropy approximation of the covariate distribution (Cressie et al., 2004). Finally, note that additional group level covariates  $Z_i$  can be incorporated multiplicatively in (5), which then becomes:  $\lambda_{0ij} \exp(\beta X_{ij}) \exp(\gamma Z_i)$  with corresponding multiplicative term in (7).

### 3.3 Logit link with binary and continuous covariate (normally distributed)

We now consider a more realistic situation where a mix of binary and continuous exposures are of interest, a case that has been considered in detail by Jackson et al. (2005). For simplicity, we shall restrict our discussion to the case where the health outcome of individual  $j$  in group  $i$  is influenced by one binary variable,  $X_{ij}^{(1)}$ , and one continuous variable  $X_{ij}^{(2)}$ , but extension to several dichotomous covariates is straightforward. In our case:

$$g(\lambda_{0ij}, X_{ij}^{(1)}, X_{ij}^{(2)}, \alpha, \beta) = \text{expit} \left( \lambda_{0ij} + \alpha X_{ij}^{(1)} + \beta X_{ij}^{(2)} \right). \quad (8)$$

Here, we have used the natural logit link to model the dependence between  $p_{ij}$  and the covariates, with  $\text{expit}(x) = \exp(x)/(1 + \exp(x))$ . In contrast to 3.1 and 3.2, no constraints are necessary. Note that in the formulation (8), we have not included any contextual variables that could explain the between group variability of the baseline  $\lambda_{0ij}$ . The logistic formulation could be extended to include these.

To obtain an explicit expression for  $p_i$ , we perform the integral (1) over the distribution of each exposure and the baseline risk using the link specified in (8). We assume that the probability of binary exposure within area  $i$  is a constant  $\phi_i$  and that the two exposures are independent. By summing over the two unknown values  $x = 0, 1$  of the binary exposure, we obtain:

$$p_i = q_{0i}(1 - \phi_i) + q_{1i}\phi_i \quad (9)$$

where  $q_{0i}$  is the marginal probability of outcome for an individual who is not exposed to the binary covariate and similarly,  $q_{1i}$  is the equivalent for an exposed individual, both given by:

$$q_{0i} = \int \text{expit}(\lambda_0 + \beta x) f_i(\lambda_0) f_i(x) d\lambda_0 dx \quad (10)$$

$$q_{1i} = \int \text{expit}(\lambda_0 + \alpha + \beta x) f_i(x) f_i(\lambda_0) d\lambda_0 dx. \quad (11)$$

Further assumptions are necessary to proceed. By reference to the discussion of (7), if both the baseline risk and the continuous covariate are constant within area, with values  $\Lambda_{0i}$  and  $m_i$  respectively, we obtain:

$$q_{0i} = \text{expit}(\Lambda_{0i} + \beta m_i) \quad (12)$$

$$q_{1i} = \text{expit}(\Lambda_{0i} + \alpha + \beta m_i). \quad (13)$$

These equations are often used even when the covariate is not constant, leading to an aggregate model that is misspecified. We refer to this model as *the naïve ecological model*. In general, as the covariate is not constant within each area, we need to integrate over its distribution  $f_i$ . By reference to Section 2.2, we assume that the baseline is constant within area,  $\Lambda_{0i}$ , and that  $X_{ij}^{(2)} \sim N(m_i, s_i^2)$ . We use a probit approximation to the logit link function to obtain instead of (12) and (13) (Jackson et al, 2005):

$$q_{0i} = \text{expit} \left\{ (1 + c^2 \beta^2 s_i^2)^{-1/2} (\Lambda_{0i} + \beta m_i) \right\} \quad (14)$$

$$q_{1i} = \text{expit} \left\{ (1 + c^2 \beta^2 s_i^2)^{-1/2} (\Lambda_{0i} + \alpha + \beta m_i) \right\} \quad (15)$$

where  $c = 16\sqrt{3}/(15\pi)$ . The use of (14) and (15) corresponds to a well specified ecological model, in contrast to the naïve model of (12) and (13).

### 3.4 Modelling of the baseline risk

The baseline risk of disease typically depends on the age and sex of the individual, and possibly other factors such as genetic susceptibility. The baseline risk will therefore vary from individual to individual, a concept that is often termed ‘frailty’ in survival analysis. Baseline risk is also likely to vary from group to group. For example, groups defined by small areas are likely to share similar lifestyle and unobserved environmental or contextual factors that may lead to between-group variations in  $\Lambda_{0i}$ .

The simplest form of between-group variability is to suppose that the  $\Lambda_{0i}$  are *exchangeable* between the groups, in other words that all the  $\Lambda_{0i}$  come from a common distribution and are independent:

$$\Lambda_{0i} \sim p(\phi_\lambda), \text{ independently for } i = 1, \dots, I.$$

Examples of commonly adopted parametric forms for  $p(\phi_\lambda)$  are a Gaussian or a Student *t*-distribution with a chosen small degree of freedom. In both cases,  $\phi_\lambda$  consists of a mean and a variance parameter.

The simple exchangeable structure is appropriate if there is no reason to suspect that some of the groups are more closely related than others in terms of their baseline risk. In many ecological designs though, the structure of the group renders this assumption implausible, in particular if the groups represent temporal or spatial units. A spatial (or temporal) structure for the  $\Lambda_{0i}$  can then be built by using a decomposition  $\Lambda_{0i} = \lambda_0 + \nu_i$ , where  $\nu_i$  is a spatially structured process, constrained to have zero mean over the areas. Typically, the specification of  $\nu_i$  will be based on a definition of ‘closeness’ between areas encoded by a neighbourhood structure: commonly,  $i \sim i'$  when the areas  $i$  and  $i'$  are contiguous. Then, a frequently-used model of spatial dependence, referred to as an intrinsic or conditional autoregressive (CAR) model, specifies the distribution of  $\nu_i$  by:

$$p(\nu_i | \nu_{i'}, i' \neq i) \sim N(\bar{\nu}_i, \sigma^2/n_i)$$

where  $\sigma^2$  is an unknown variance parameter and  $\bar{\nu}_i = \sum_{s: i' \sim i} \nu_{i'}/n_i$ ,  $n_i$  denotes the number of neighbours of area  $i$ , and we impose that  $\sum \nu_i = 0$ . The CAR model has been extensively used in disease mapping studies concerned with rare diseases after its introduction by Besag et al (1991). The resulting estimates of the  $\{\Lambda_{0i}\}$  borrow strength from the neighbouring areas and are smoothed towards a local mean.

## 4 Combining aggregated and individual level data

Once the aggregated model has been derived from individual level specification, it is reasonably straightforward to combine it with a sample of individual level information. For example, the ecological regression model defined by the binomial distribution (2) together with (14) and (15), can

be combined with (8) (where the base line is  $\Lambda_{0i}$ ). The models will have in common the parameters  $\alpha$ ,  $\beta$  and  $\Lambda_{0i}$ . The ecological data will consist of values for  $y_i$ ,  $N_i$ , the proportion  $\bar{X}_i^{(1)}$  of people with dichotomous covariate taking the value 1, and values for  $m_i$  and  $s_i^2$  (the model could be easily extended to include instead a sample of values of the continuous covariate modelled as  $X_{ij}^{(2)} \sim N(m_i, s_i^2)$ ). The individual data will consist of a sample of values of  $y_{ij}$ ,  $X_{ij}^{(1)}$ ,  $X_{ij}^{(2)}$ . Note that it is not necessary to have individual data available on all the units for which there are aggregated data.

As discussed previously, different models can be used to borrow information across the areas to estimate the baseline risks  $\Lambda_{0i}$ . In this particular case, we suppose that the  $\Lambda_{0i}$  are exchangeable and follow a normal distribution. Once the individual and ecological regression have been specified, a Bayesian hierarchical model is then used to carry out the data integration. Care has to be taken when choosing the priors for  $\alpha$ ,  $\beta$  and for the mean and variance of the exchangeable distribution for  $\log \Lambda_{0i}$ , as standard flat priors would lead to marginal priors for  $p_{ij}$  that are biased toward 0 or 1. Normal priors for  $\alpha$ ,  $\beta$  with suitable variances, a logistic prior for the mean of the random effect and an inverse Gamma for the variance were chosen (see Jackson et al., 2005 for full details).

Some of the benefits of integrating ecological and individual level data are illustrated in Table 1, which reports part of the results of a comprehensive simulation study. The results correspond to averages over 50 simulated data sets with  $I = 100$  groups of  $N_i = 1000$  individuals,  $\alpha = \log(2) = 0.69$ ,  $\beta = \log(2.3) = 0.8$ , a 95% sampling interval for the baseline disease risk  $\Lambda_{0i}$  of (0.07, 0.14) and values for  $\bar{X}_i^{(1)}$  equally spaced between 0 and 0.2 (this narrow range is challenging for the analysis but realistic if for example,  $X_i^{(1)}$  represents smoking). The simulated data for  $X_{ij}^{(2)}$  are illustrated in Figure 2, which represents the within-area distribution of  $X_{ij}^{(2)}$  for each of the 100 areas. Note that there is a small negative correlation (-0.3) between  $m_i$  and  $s_i^2$ . Within each group  $i$ ,  $X_{ij}^{(2)}$  are fixed to be 100 equally spaced quantiles of  $N(m_i, s_i^2)$ .

Altogether, we see in Table 1 that the risk estimate  $\beta$  related to the continuous covariate is well estimated, even with a naïvely specified model. This may well reflect the fact that the between-area variation in mean exposure to the continuous covariate is reasonably large compared with the within area variance, and so the ecological data contains a reasonable amount of information. However, in view of the narrow range of proportion exposed to the dichotomous covariate, there is little information in the ecological data to estimate it. We see that the risk estimate  $\alpha$  is indeed estimated altogether with wider uncertainty. When using the naïve ecological model, it is clearly underestimated and there is a marked improvement when the information of the individual sample is integrated, in particular with the naïve ecological model. Individual data alone leads to risk estimates with

TABLE 1. Posterior mean and posterior SD of the risk estimates, (average over 50 replicates). True values  $\alpha = 0.69$   $\beta = 0.8$ .

Model	Sample	$\alpha$ (SD)	$\beta$ (SD)
Ecological (naïve)	0	0.58 (0.32)	0.78 (0.09)
Combined (naïve)	10	0.66 (0.24)	0.79 (0.09)
Ecological (well specified)	0	0.62 (0.33)	0.81 (0.10)
Combined (well specified)	10	0.65 (0.24)	0.81 (0.10)
Individual data only	10	0.64 (0.29)	0.81 (0.22)

sizeable intervals, while the combined models (naïve and well specified) give both narrower intervals and small bias.

Other cases were simulated corresponding to the more complex situation of correlated exposures (i.e. a different within group distribution of the continuous covariate following the value of the dichotomous one) or an additional interaction term for the effect of both exposures on  $p_{ij}$ . In each case, even with appropriate modification of the specification of the aggregated model, the information in the ecological data alone is limited and combining with a sample of individual level values reduces the bias (see Jackson et al., 2005).

## 5 Ecological Analysis of the risk factors for Limiting Long-Term Illness

To illustrate some of the ideas discussed above in a real application, we carry out an ecological study of the prevalence of limiting long-term illness (LLTI) among men aged between 40 and 59 years of age in London, UK. It was the only health outcome systematically recorded in the 1991 census, using the question “Does the person have any long-term illness, health problem or handicap which limits his/her daily activities or the work he/she can do. Include problems which are due to old age”. Its interpretation has been discussed by Cohen et al. (1995), who found it to be correlated with several illnesses such as arthritis, asthma, chronic bronchitis, heart disease or diabetes. Note that here we restrict our attention to middle aged men. Aggregate data on limiting long-term illness, age, sex and ethnicity for all 761 electoral wards in London are taken from the UK census in 1991. This excludes some wards with very small resident population. From the census, characterisation of the socio-economic deprivation of each ward can also be obtained. We have chosen to use the classical Carstairs deprivation index (Carstairs and Morris, 1991) based on rates of male unemployment, car ownership, low social class and household overcrowding. Estimates of the mean household income, in UK pounds, for each ward are obtained from the ACORN consumer classification database (CACI, Limited). We also have small samples of individual-level data available from the Health Survey

for England (HSE) (Department of Health, U.K., <http://www.dh.gov.uk>), including limiting long-term illness, age, sex, ethnicity, and income. Individual ward identifiers were made available under a special arrangement with the data providers. Individual-level data are available from 255 wards, with 1–9 observations per ward (median 1.6). Thus, in this case the individual data is very sparse and does not cover all the wards for which the aggregated data are available. We are interested in characterising the effect of ethnicity and income on LLTI, and to investigate whether there is a residual contextual effect of area-level deprivation.

The basic individual logistic model that we consider is given by:

$$\text{logit}(p_{ij}) = \Lambda_{0i} + \gamma Z_i + \alpha X_{ij}^{(1)} + \beta X_{ij}^{(2)}. \quad (16)$$

where  $Z_i$  characterises the deprivation of the ward,  $X_{ij}^{(1)}$  is ethnicity (dichotomous white/non white) and  $X_{ij}^{(2)}$  is log income. Since the analysis is restricted to one sex and age class, it seems reasonable to assume that the baseline is the same for all individuals in the group. The variability of  $\Lambda_{0i}$  quantifies the remaining contextual or environmental sources of heterogeneity of LLTI prevalence. The naïve, well specified ecological models and the combined models are derived along the lines of Section 4. Exchangeable normal random effects were used for  $\Lambda_{0i}$ . One additional modification to the combined model was rendered necessary in view of the discrepancy between the overall prevalence of LLTI as reported by the census and the HSE: 15% and 26% respectively among men aged 40 to 59 years. Cohen et al. (1995) found a similar discrepancy between census and survey data on LLTI in Scotland. A constant increment of 0.7 is thus added to the logit baseline of the individual-level component in the combined model.

An exploratory analysis of the between-group variability of the aggregate covariates indicate a somewhat similar pattern between the wards, with wards with a higher proportion of non white tending to have low average income, both covariates being also linked to the prevalence of LLTI, and contextual variables such as the deprivation index, (see Figure 1). Log household income has a wide within-area variability compared to its between-area variability (Figure 2), although the variance is approximately constant between areas, which may prevent specification bias when within-area variability is ignored, as discussed in Section 3.2.

Results of the fit of a variety of models are summarised in Table 2. The individual analysis indicates a negative effect of non-white ethnicity and a negative effect of income on the risk of LLTI (first line of Table 2), but the power is low and the interval estimates are wide and inconclusive for the effect of non-white ethnicity. The coefficients  $\alpha$  and  $\beta$  are not changed when area-level deprivation is included in the analysis of the individual data, but this is to be expected as the data are very sparse in each ward. On the other hand, at the aggregate level, the inclusion of deprivation has a marked influence on the estimates of the coefficients  $\alpha$  and  $\beta$ , indicating substantial

TABLE 2. Estimated coefficients of non-white ethnicity and log income, considered as individual effects, with or without additional contextual effect of deprivation, for individual level, ecological (naïve and well specified), combined regression models, and combined regression with spatially-correlated instead of exchangeable baseline risks.  $SD(\Lambda_{0i})$  is the estimated standard deviation parameter of the exchangeable random effects, or the empirical standard deviation of the spatially-correlated random effects.

Model	Non-white ethnicity	Log income	Deprivation	$SD(\Lambda_{0i})$
Individual	-0.29 (-0.88, 0.28)	-0.56 (-0.80, -0.33)	–	0.17 (0.053, 0.56)
Individual	-0.36 (-0.98, 0.23)	-0.55 (-0.80, -0.32)	-0.022(-0.032,0.074)	0.18 (0.052, 0.64)
Ecological (naïve)	1.02 (0.88, 1.16)	-1.35 (-1.45, -1.25)	–	0.24 (0.23, 0.26)
Ecological (naïve)	0.27 (0.15, 0.39)	-0.57 (-0.67, -0.47)	0.068 (0.063, 0.074)	0.17 (0.16, 0.18)
Ecological (well specified)	1.38 (1.22, 1.53)	-2.13 (-2.35, -1.90)	–	0.33 (0.30, 0.35)
Ecological (well specified)	0.35 (0.23, 0.47)	-0.71 (-0.84, -0.60)	0.066 (0.061, 0.072)	0.18 (0.16, 0.19)
Combined (well specified)	1.35 (1.21, 1.51)	-2.05 (-2.27, -1.83)	–	0.32 (0.29, 0.34)
Combined (well specified)	0.34 (0.20, 0.47)	-0.71 (-0.85, -0.57)	0.067 (0.061, 0.073)	0.18 (0.16, 0.19)
Combined (spatial)	1.00 (0.85, 1.15)	-1.75 (-1.87, -1.64)	–	0.30 (0.29, 0.31)
Combined (spatial)	0.26 (0.12, 0.44)	-0.86 (-1.04, -0.73)	0.060 (0.050, 0.067)	0.19 (0.17, 0.20)

unmeasured confounding if contextual variables are not included. Once deprivation is included in the model of the baseline, the estimates for  $\alpha$  and  $\beta$  become more compatible between the individual and well specified ecological model, in the sense that interval estimates overlap. The combined model estimates are little changed from the well specified ecological model in this case due to the sparsity of the individual level data. Non-white ethnicity and income are clearly associated with LLTI, in opposite directions. Future work in this area should address design issues, to maximise the amount of information available in the individual-level data.

We were further interested in characterising the pattern of the residual baseline rate across the wards. Figure 3 shows a map of these residual risks for London wards, for the combined model of Table 2, after adjusting for area-level deprivation. A spatial pattern appears quite clearly, suggesting that fitting spatial random effects as described in Section 3.4 rather than an exchangeable normal model for the baseline is appropriate. When this spatial correlation is modelled, the coefficients of ethnicity and log income are moderately altered but association stays significant. The estimated spatially-correlated baseline rates are represented in the map of Figure 4. The baseline pattern is smoothed with comparison to Figure 3. The posterior mean Deviance Information Criterion (Spiegelhalter et al., 2003) of the spatial model is 5915, compared to 5982 for the model with exchangeable baselines, indicating that the spatial model gives a significant improvement in fit.

**Acknowledgments:** This work was supported by the Economic and Social Research Council, award number R000239598. SR and NB acknowledge

FIGURE 1. Relationships between ward-level census variables

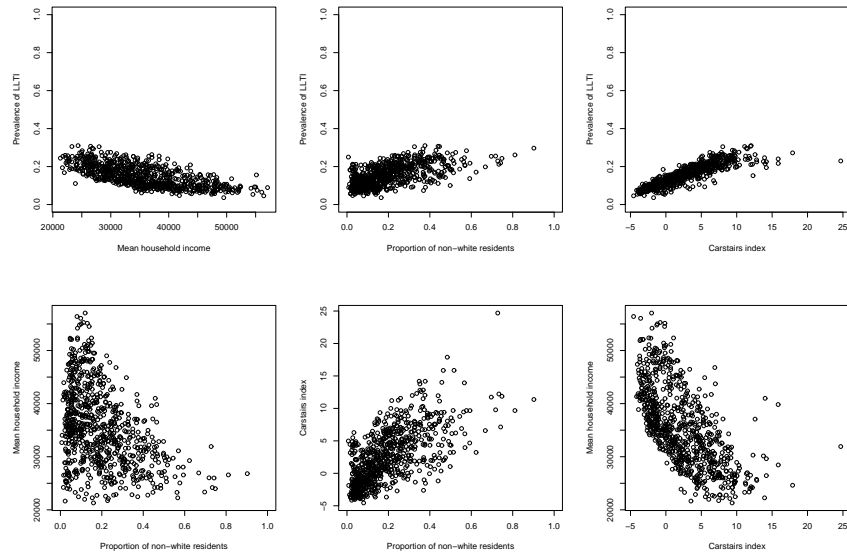


FIGURE 2. Left: Simulated continuous covariate data. Mean and 2.5%–97.5% quantiles of  $X_{ij}^{(2)}$  for 100 areas. The mean within-area standard deviation is 0.38, the standard deviation of within-area means is 0.28. Right: distribution of log household income for 50 randomly-chosen London wards out of 761. Mean  $\pm 1.96 \times$  standard deviation of log income.

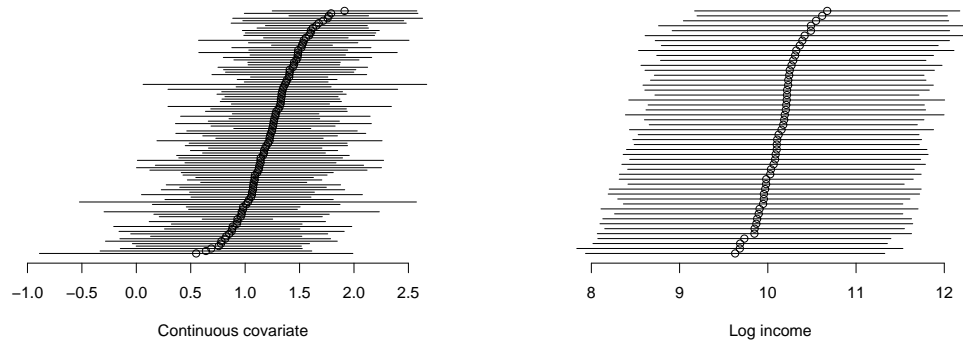


FIGURE 3. Map of residual risks of LLTI across London wards,  $\lambda_{0i}$ , modelled as exchangeable random effects.

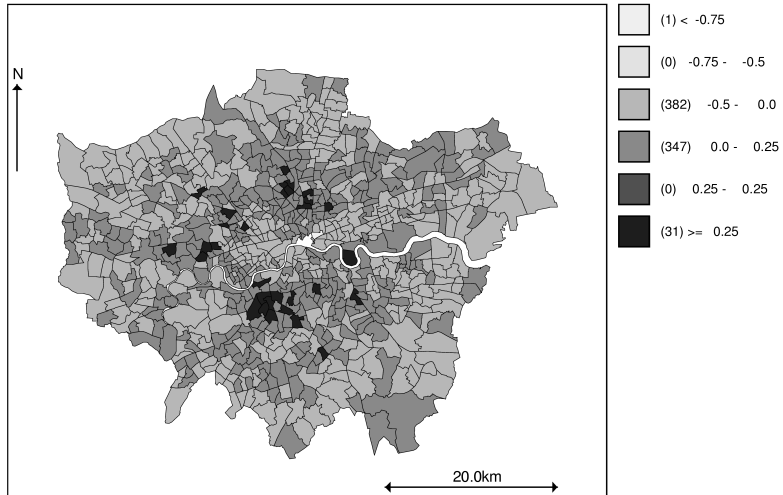
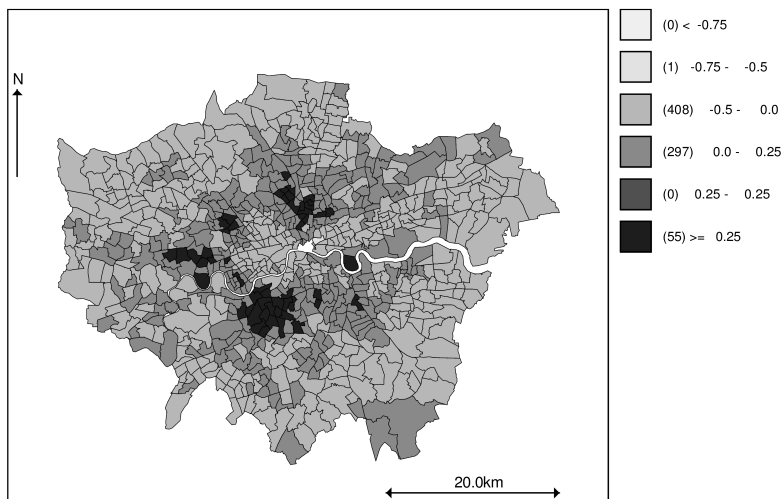


FIGURE 4. Map of residual risks of LLTI across London wards,  $\lambda_{0i}$ , modelled as spatially-correlated random effects.



partial support from AFSSE RD2004004 and INSERM-ATC A03150LS French grants. Thanks to the Small Area Health Statistics Unit, CACI Limited, the UK Census Dissemination Unit and the Health Survey for England for provision of data.

## References

- Ben-Shlomo, Y., White, I. and Marmot, M. (1996). Does the variation in socioeconomic characteristics of an area affect mortality? *British Medical Journal*, **312**, 1013–4.
- Besag, J., York, J., Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.
- Best, N., 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: Statistics In Society*, **164**(1), 155–174.
- Carstairs, V. and Morris, R. (1991). Deprivation and health in Scotland. Aberdeen University Press, Aberdeen.
- Cohen, G., Forbes, J., and Garraway, M. (1995). Interpreting self-reported limiting long-term illness. *British Medical Journal*, **311**(7007), 722–24.
- Cressie, N., Richardson, S. and Jaussent, I. (2004). Ecological bias: use of maximum- entropy approximations. *Australian and New Zealand Journal of Statistics*, **46**, 233–255.
- Elliott, P.E., Wakefield, J., Best, N.G. and Briggs, D.J. (2000). *Spatial Epidemiology*. Oxford University Press, Oxford.
- Greenland, S. and Morgenstern, H. (1989). Ecological bias, confounding and effect modification. *International Journal of Epidemiology*, **18**, 269–284.
- Greenland, S. and Robins, J. (1994). Ecological studies — biases, misconceptions and counterexamples. *American Journal of Epidemiology*, **39**, 47–760.
- Greenland, S. (2003). Divergent biases in ecologic and individual-level studies. *Statistics in Medicine*, **11**, 1209–23.
- Jackson, C., Best, N. and Richardson, S. (2005). Improving ecological inference using individual-level data. *Technical report*. Imperial College. [www.bias-project.org.uk](http://www.bias-project.org.uk)

- King, G. (1997). *A solution to the ecological inference problem*. Princeton. Princeton University Press.
- Lasserre, V., Guihenneuc-Jouyaux, C. and Richardson, S. (2000). Biases in ecological studies: utility of including within-area distribution of confounders. *Statistics in Medicine*, **19**, 45–59.
- Piatadosi, 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**, 113–26.
- Richardson, S., Stucker, I. and Hémon, D. (1987). Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Epidemiology*, **16(1)**, 111–120.
- Richardson, S. (1992). Statistical methods for geographical correlation studies. In *Geographical and environment epidemiology: methods for small area studies*. 181–204. Elliott, P.E., Cuzick, J., English, D., and Stern, R. eds. Oxford University Press, Oxford.
- Richardson, S. and Monfort, C. (2000). Ecological correlation studies. In *Spatial Epidemiology*, chapter 11. Elliott, P.E., Wakefield, J., Best, N.G. and Briggs, D.J. eds, Oxford University Press, Oxford.
- Richardson, S., and Best, N. (2003). Bayesian hierarchical models in ecological studies of health-environment effects. *Environmetrics*, **14**, 129–147.
- Sheppard, L. (2003). Bias and information in group-level studies. *Biostatistics*, **4(2)**, 265–278.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society, Series B*, **64(4)**, 583–639.
- Wakefield, J. and Salway, R. (2001). A statistical framework for ecological and aggregate studies. *Journal of The Royal Statistical Society, Series A: Statistics In Society*, **164(1)**, 119–137.
- Wakefield, J. (2003). Sensitivity analyses for ecological regression. *Biometrics*, **59**, 9–17.
- Wakefield, J. (2004). Ecological inference for  $2 \times 2$  tables (with discussion). *Journal of The Royal Statistical Society, Series A: Statistics In Society*, **167(3)**, 385–445.