UW Biostatistics Working Paper Series

12-20-2005

A Hybrid Model for Reducing Ecological Bias

Ruth Salway
University of Bath, UK, R.E.Salway@bath.ac.uk

Jon Wakefield
University of Washington, jonno@u.washington.edu

A Hybrid Model for Reducing Ecological Bias

Ruth Salway (corresponding author)

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

R.E.Salway@bath.ac.uk

Tel: +44 1225 386320 Fax:+44 1225 386492

Jon Wakefield

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

December 20, 2005

Abstract

A major drawback of epidemiological ecological studies, in which the association between area-level summaries of risk and exposure are used to make inference about individual risk, is the difficulty in characterising within-area variability in exposure and confounder variables. To avoid ecological bias, samples of individual exposure/confounder data within each area are required. Unfortunately these may be difficult or expensive to obtain, particularly if large samples are required. In this paper we propose a new approach suitable for use with small samples. We combine a Bayesian non-parametric Dirichlet process prior with an estimating functions approach, and show that this model gives a compromise between two previously-described methods. The method is investigated using simulated data, and a practical illustration is provided through an analysis of mortality and income data across

England. We conclude that we require good quality prior information about the exposure/confounder distributions and a large between- to within-area variability ratio for an ecological study to be feasible using only small samples of individual data.

Keywords: aggregate data; Dirichlet process prior; ecological fallacy; pure specification bias; within-area variability.

1 Introduction

Many disciplines make use of aggregate data, including epidemiology, social sciences and education; Salway and Wakefield (2004) provide a comparison between models and approaches in epidemiology and the social sciences. In environmental and social epidemiology aggregate data may consist of area-level disease rates and summary exposure and confounder values within each area. Usually, the purpose of such ecological, sometimes called geographical correlation, studies is to make inference at the individual level. The major problem with such a study design is the potential for ecological bias, which is due to aggregation. Many authors have documented the sources of ecological bias, see for example Greenland and Robins (1994) and Richardson and Montfort (2000), but far fewer have proposed solutions.

The term *ecological bias* is generally used to describe bias that may arise from several different sources in aggregate data (Greenland and Morgenstern, 1989). We will concentrate on bias that arises when aggregating a nonlinear individual-level model over the within-area distribution of covariates. Such bias is caused by within-area variability in both the exposure of interest and in confounders; this will be referred to as *within-area variability bias* (Greenland, 1992, uses the term *pure specification bias*).

Existing approaches to correct for within-area variability bias require individual data on exposures in each area. Since we do not require the individual link between covariate data and health

outcome, it is feasible for mortality or morbidity data to be taken from one source, and individual data from another. For example, disease counts may be obtained from a cancer registry, and individual data from a survey. One of the advantages of ecological studies is in situations where collecting individual data is either difficult or expensive. In these situations it may be possible to obtain only very small samples of individual data. Unfortunately, existing approaches require larger samples; simulations (Wakefield and Salway, 2001) have suggested samples of at least 100 in each area are required.

The aim of this paper is to describe a new model that reduces within-area variability bias when only small samples of covariate data are available. The method incorporates prior information about the data, which will be crucial when only very small samples of individual data are available.

The paper is organised as follows. In Section 2 we provide a motivating example concerning the association between mortality and income, and describe the data we will use subsequently. In Section 3 we look in detail at what causes within-area variability bias in order to understand how it arises and how it may be removed. In particular we consider how existing approaches perform when we have only small samples of individual data. In Section 4 we describe our new method. In Section 5 a series of simulations compares the new model with existing methods across a range of scenarios. Section 6 presents a practical example using the income data, and Section 7 provides a concluding discussion.

2 Motivating Example: Income and Health

To motivate the work that follows we take an example from social epidemiology; an analysis of the relationship between ill-health and income using ecological data. An extensive literature suggests that mortality rates decrease as average income increases (see for example Judge et al.,

1998). Current debate has centred around two hypotheses: the *absolute income* hypothesis, where individual health is affected by individual income, and the *relative income* hypothesis, where health depends also on the degree of income inequality in neighbourhood of residence. While aggregate data have been used to investigate these hypotheses, their usefulness is debated (Gravelle et al., 2002); in part, this is due to the problems of interpretation in the presence of ecological bias.

We will illustrate the methods described in this paper using publicly available data from the UK Data Archive (`www.data-archive.ac.uk`). The study population is males aged over 64 years in England, and the areas consist of 28 Strategic Health Authorities (SHA), with total population sizes between 60,700 and 198,100. The response is all-cause mortality in each area for 2002 (UK Data Archive: SN4817, 2002). The risk factor of interest is *equivalised household income*, based on the McClements score (McClements, 1977); this score adjusts total income to take account of the size and composition of the household, so that for example, a family with three children is considered to be less well off than a single person living alone on the same income. The Health Survey for England 2002, provides equivalised incomes on individual samples of data, with sample sizes between 3 and 44 in each SHA. The data are taken from different sources, and so an individual study is not possible as the link between income and mortality is not available. The sample sizes in this case are very small; for comparison we also analysed the same data for all males, giving sample sizes between 60 and 216. This is a fairly crude analysis, and is for illustrative purposes only; it is not our intention to demonstrate support for either income hypothesis.

We consider the relationship between mortality and the logarithm of income, taken to base 2 and use a log-linear model. Using logarithms to base 2 within this model means that, causally interpreted, the effect of a doubling of income is constant; so increasing an annual income from £5,000 to £10,000 has the same effect on mortality as increasing an income from £20,000 to

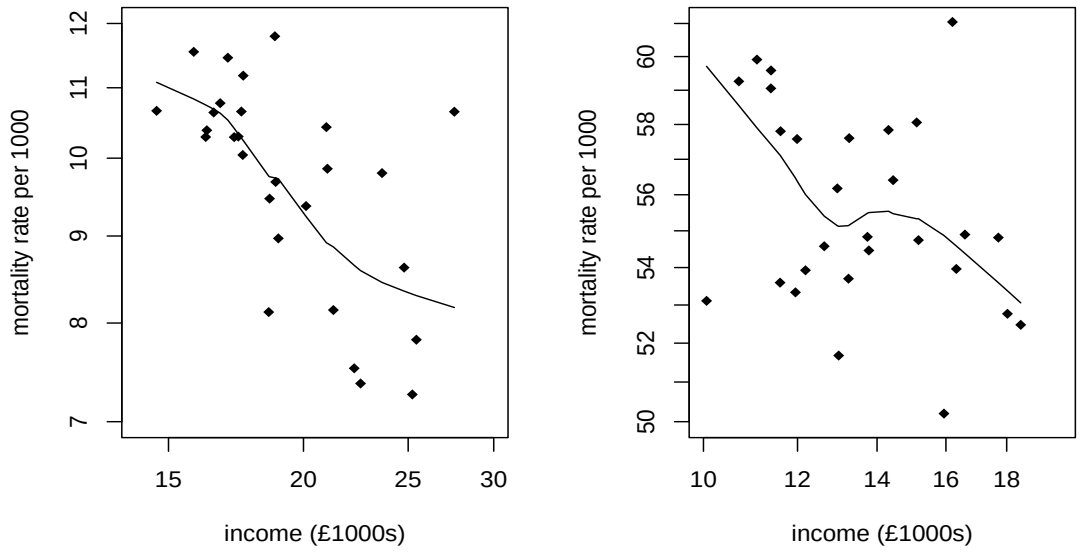

Figure 1: Log mortality rate against mean log income (in thousands of pounds), with smoother superimposed, for men of all ages (left) and men over 64 (right).

£40,000. This is a more realistic model than using untransformed income, which would assume the same effect of going from £5,000 to £10,000 as from £95,000 to £100,000.

The majority of the variability in income is between areas. Figure 1 shows plots of the log mortality rate against mean log income; the ecological data suggest a negative relationship between mortality and income for both age populations. Figure 2 illustrates the degree of within-area variability in logged income; there is considerable within-area variability, which indicates the potential for considerable ecological bias.

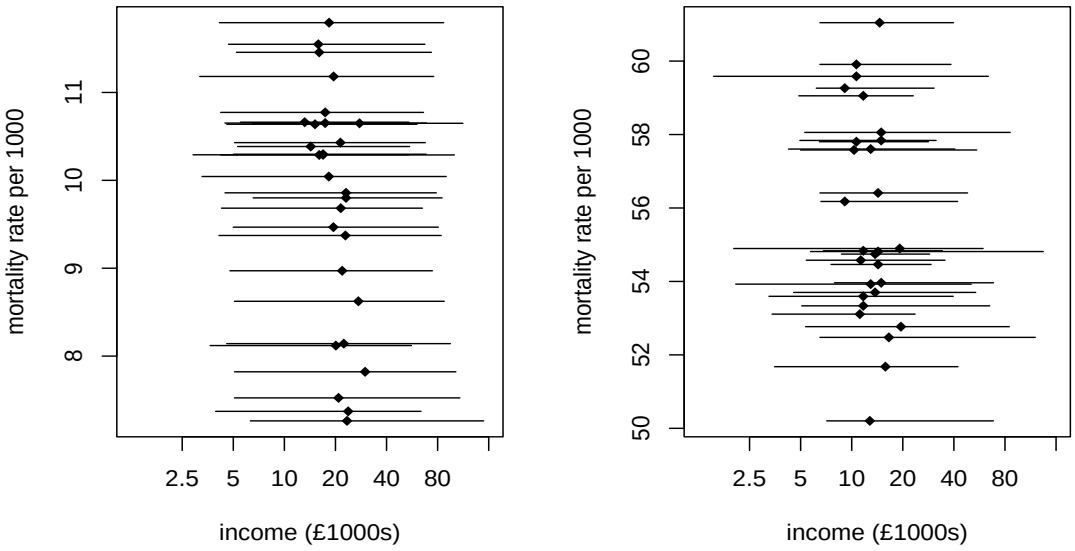

Figure 2: Within-area variability in log income: median estimates with 95% intervals based on percentiles for men of all ages (left) and men over 64 (right).

3 Ecological Models and Ecological Bias

3.1 Notation

Consider a study area partitioned into a disjoint set of K areas, each containing n_k individuals, with $k = 1, \ldots, K$. Using terminology from epidemiology, let Y_{ki} be a Bernoulli random variable representing the disease outcome of individual i in area k, over a specific time period, with $Y_{ki} = 1$ representing a case, and $Y_{ki} = 0$ a non-case, $i = 1, \ldots, n_k$. We are interested in how this outcome is related to an exposure variable of interest, X_{ki}.

We begin by specifying the disease/covariate relationship for an individual (following Richardson et al., 1987; Prentice and Sheppard, 1995; Wakefield and Salway, 2001), which can be thought of as the model that we would fit if individual data $\{Y_{ki}, X_{ki}\}$ were available. This approach emphasises that we are interested in estimates of the individual effect. Since most diseases may be considered rare in a statistical sense, a common individual-level model in epidemiology is:

$$\begin{aligned} Y_{ki}|\beta_0, \beta_1, X_{ki} &\sim_{\text{ind}} \text{Bern}\left\{p_I(\beta_0, \beta_1, X_{ki})\right\}, \\ p_I(\beta_0, \beta_1, X_{ki}) &= \exp\left\{\beta_0 + \beta_1 X_{ki}\right\} \end{aligned} \tag{3.1}$$

where the subscript I emphasises that $p_I(\cdot)$ characterises the individual relationship. In (3.1), $\exp(\beta_0)$ is the baseline risk and $\exp(\beta_1)$ is the relative risk corresponding to an increase of one unit in the variable of interest.

Model (3.1) is simple and does not take into account other possible causes of ecological bias, such as confounding or contextual effects, and in addition we have considered only a single continuous variable. However, it is straightforward to extend the individual formulation to multivariate exposures and confounders.

In an ecological study, we typically only have total disease counts, $Y_k = \sum_{i=1}^{n_k} Y_{ki}$, and some

summary of the exposure distribution, X_k. In this paper, we assume also that we have a sample of individual covariate data of size m_k with $2 \leq m_k \leq n_k$, in each area k and we denote these data by $\mathbf{X}_k^{m_k} = \{X_{kj} : j = 1, \ldots m_k\}$. We will assume that X_k is the mean of the sample, that is $X_k = \sum_{j=1}^{m_k} \mathbf{X}_k^{m_k}/m_k$.

3.2 Ecological Bias

An obvious choice of ecological model is the individual model (3.1) with individual data replaced by ecological data, to give the *simple ecological* model

$$\begin{aligned} Y_k|\beta_0^*, \beta_1^*, X_k &\sim_{\text{ind}} \text{Po}\left\{n_k p_I(\beta_0^*, \beta_1^*, X_k)\right\}, \\ p_I(\beta_0^*, \beta_1^*, X_k) &= \exp\left\{\beta_0^* + \beta_1^* X_k\right\}. \end{aligned} \tag{3.1}$$

where β_1^* is the ecological effect parameter. Ecological bias arises when the simple ecological model (3.1) does not estimate the same parameters as the individual model (3.1); that is $\beta_1^* \neq \beta_1$. Model $p_I(\cdot)$ is a convex function, and so bias will occur since, by Jensen's inequality, we have

$$E\left[p_I(\beta_0, \beta_1, X_{ki})\right] \leq p_I\left(\beta_0, \beta_1, \mu_k\right), \tag{3.2}$$

where the expectation is over the within-area exposure distribution, and the right hand side is the simple ecological model (3.1) evaluated at the true area mean μ_k. Equality will occur if and only if $X_{ki} = X_k$ for all i, that is, if there is no within-area variation in individual exposures. There is also no bias if $p_I(\cdot)$ is linear in X, and so within-area variability bias arises only for a nonlinear model.

In specific circumstances we may still have no bias. We write

$$E[Y_k|\boldsymbol{\phi}_k] = n_k e^{\beta_0} \int_x f_k(x|\boldsymbol{\phi}_k) e^{\beta_1 x} \mathrm{d}x, \tag{3.3}$$

where $f_k(x|\boldsymbol{\phi}_k)$ is the within-area distribution of X in area k, with parameters $\boldsymbol{\phi}_k$. Expanding

the $e^{\beta_1 x}$ term in a Taylor series about μ_k gives

$$E[Y_k|\boldsymbol{\phi}_k] = n_k \exp(\beta_0 + \beta_1 \mu_k) \sum_{r=0}^{\infty} \frac{\beta_1^r}{r!} \mu_k^{(r)} \tag{3.4}$$

where $\mu_k^{(r)} = E\left\{(X_{ki} - \mu_k)^r\right\}$ is the rth central moment of the within-area exposure distribution $f_k(\cdot)$. The summation term above is exactly the bias component given by Richardson et al. (1987) and consists of terms involving higher moments of the within-area exposure distribution. There will be no bias whenever the summation term is independent of the mean, irrespective of the distribution of X; this occurs when the second and higher moments do not depend on μ_k. In general, of course, this will not be the case; this result is of more mathematical than practical interest. However, it does suggest that the mean-variance relationship, and in particular the strength of that relationship, is an important factor in the size of ecological bias since this is the dominant term in (3.4) for $\beta_1 < 1$. (this corresponds to a relative risk of less than 2.7, which is typical in studies of environmental pollutants; see the examples in Wakefield, 2003). Higher moments become increasingly small and so contribute little to the bias.

3.3 Ecological Models

One obvious solution to ecological bias is to consider explicitly the model obtained by aggregating over the within-area exposure distribution. If all exposures within an area are assumed independent, then

$$\begin{aligned} Y_k|\boldsymbol{\phi}_k &\sim_{\text{ind}} \text{Po}\{n_k p_E(\beta_0, \beta_1, \boldsymbol{\phi}_k)\} \\ p_E(\beta_0, \beta_1, \boldsymbol{\phi}_k) &= n_k \exp\left\{\beta_0 + K(\beta_1)\right\} \end{aligned} \tag{3.1}$$

where $K(\beta_1)$ is the cumulant generating function of the within-area distribution (Richardson et al., 1987), and the parameters β_0, β_1 are the same as in the individual model (3.1). Expression (3.1) may be interpreted as the *average individual risk* within area k, and will generally depend on the mean of X, and on higher moments of the distribution, as seen in (3.4). If the exposures

are dependent, then $Y_k|\boldsymbol{\phi}_k$ will have the same expectation, but the distribution will no longer be Poisson.

We refer to model (3.1) as the *parametric ecological model*. A convenient assumption, and one that may often be suitable in practice, is to take the within-area distributions as approximately normal (the *parametric normal ecological model*), with $X_{ki} \sim N(\mu_k, \sigma_k^2)$, in which case (3.1) becomes

$$E[Y_k|\beta_0, \beta_1, \boldsymbol{\phi}_k] = n_k \exp(\beta_0 + \beta_1\mu_k + \beta_1^2\sigma_k^2/2) \tag{3.2}$$

In practice, we will need to use estimates of the unknown parameters μ_k and σ_k^2. While accurate estimation of the mean may often be possible, information about the within-area exposure variances σ_k^2 is unlikely to be routinely available. So in practice, a sample of exposure data $\mathbf{X}_k^{m_k}$ in each area is required to estimate σ_k^2, and, in particular when the size of the subsample is small, estimation of this variance may result in bias due to errors-in-variables.

When the within-area exposure distribution is not normal, expression (3.2) can be seen as a second-order approximation to the true model (3.1). Thus in practice (3.2) may be an adequate approximation provided that within-area exposure distributions are not heavily skewed. For very heavily skewed distributions, or larger exposure effects, expression (3.1) may sometimes be available in closed form (for example, the gamma distribution given in Wakefield and Salway, 2001). The log-normal distribution is often used to model environmental exposures (for a theoretical justification of its use, see Ott, 1994), but cannot be used here since the moment generating function does not exist. In this case the Taylor expansion (3.4) may provide a suitable approximation by including higher terms, although in general the use of higher-order moments will introduce further inaccuracies due to the increased instability in estimating higher moments such as the skewness.

Prentice and Sheppard (1995) proposed a model that makes no assumption about the within-area

distribution, but instead explicitly uses the sample of individual-level data $\mathbf{X}_k^{m_k}$ to empirically estimate (3.3). For a sample of size m_k, we have

$$E[Y_k|\beta_0, \beta_1, \mathbf{X}_k^{m_k}] = n_k \widehat{\theta}_{Ak}, \tag{3.3}$$

where

$$\widehat{\theta}_{Ak} = \frac{1}{m_k} \sum_{j=1}^{m_k} \exp(\beta_0 + \beta_1 X_{kj}) \tag{3.4}$$

is an estimate of the individual average risk, based on the sample data. Following the terminology of Prentice and Sheppard (1995) we will refer to model (3.3) as the *aggregate model*. It may be fitted using an estimating equations approach; see Prentice and Sheppard (1995) for details. When using any sample of the exposure data, $\mathbf{X}_k^{m_k}$, rather than all the data within the area, $\mathbf{X}_k^{n_k}$, the estimating equation in the aggregate approach is biased in expectation over all possible choices of sample; this introduces bias in the estimate of β_1. For large samples this is negligible, but problems arise with small samples, since the finite sampling bias in the estimating equations increases as sample sizes decrease.

Prentice and Sheppard (1995) propose an adjusted estimating equation to correct the finite sampling bias (the *corrected aggregate model*). However, this requires estimation of an additional term and they suggest that in practice the increase in variability will outweigh the benefits. Simulations have shown (Wakefield and Salway, 2001; Sheppard et al., 1996, see also the simulations in Section 3.4) that while this corrected version can perform very well for smaller samples (for example, less than 100), for very small samples, say $m_k < 20$, the estimator becomes unstable, which results in convergence problems for the estimation algorithm (we have observed problems of both non-convergence and convergence to the wrong value). As sample sizes decrease non-convergence rates can be above 50%.

3.4 Using small samples of individual data

We simulated 10 datasets with normally distributed within-area distributions for 50 areas containing 2000 individuals each. The means and variances are linearly related, with a between to within area variance ratio of around 1, and a negative effect parameter $\beta_1 = -\log(2)$ (motivated by the income example). Figure 3 illustrates the bias in the estimate of β_1 as a function of sample size. The size of bias is shown for the simple ecological model, which is always biased, the parametric normal model, and the uncorrected and corrected aggregate models. In this scenario, the ecological bias is towards the null, and bias from using small samples is also towards the null, resulting in negative bias for all models. Estimation becomes poorer for both parametric and aggregate approaches as the sample size decreases. In particular, the bias in the uncorrected aggregate and the parametric methods is extremely large for small samples of $m_k < 20$. Note also the variability in the corrected aggregate method, reflecting its increased variance, and the instability for very small sample sizes.

4 Hybrid Ecological Model

4.1 The Hybrid Ecological Model

As in (3.1), we assume the individual disease model is given by

$$\begin{aligned} Y_{ki}|\beta_0, \beta_1, X_{ki} &\sim_{\text{ind}} \text{Bern}\left\{p_I(\beta_0, \beta_1, X_{ki})\right\}, \\ p_I(\beta_0, \beta_1, X_{ki}) &= \exp\left\{\beta_0 + \beta_1 X_{ki}\right\}. \end{aligned}$$

We write $X_{ki}|F_k \sim F_k$ where F_k is the unknown distribution function of X in area k, and assume that

$$F_k|\alpha_k, F_{0k} \sim \text{DP}(\alpha_k, F_{0k}), \tag{4.1}$$

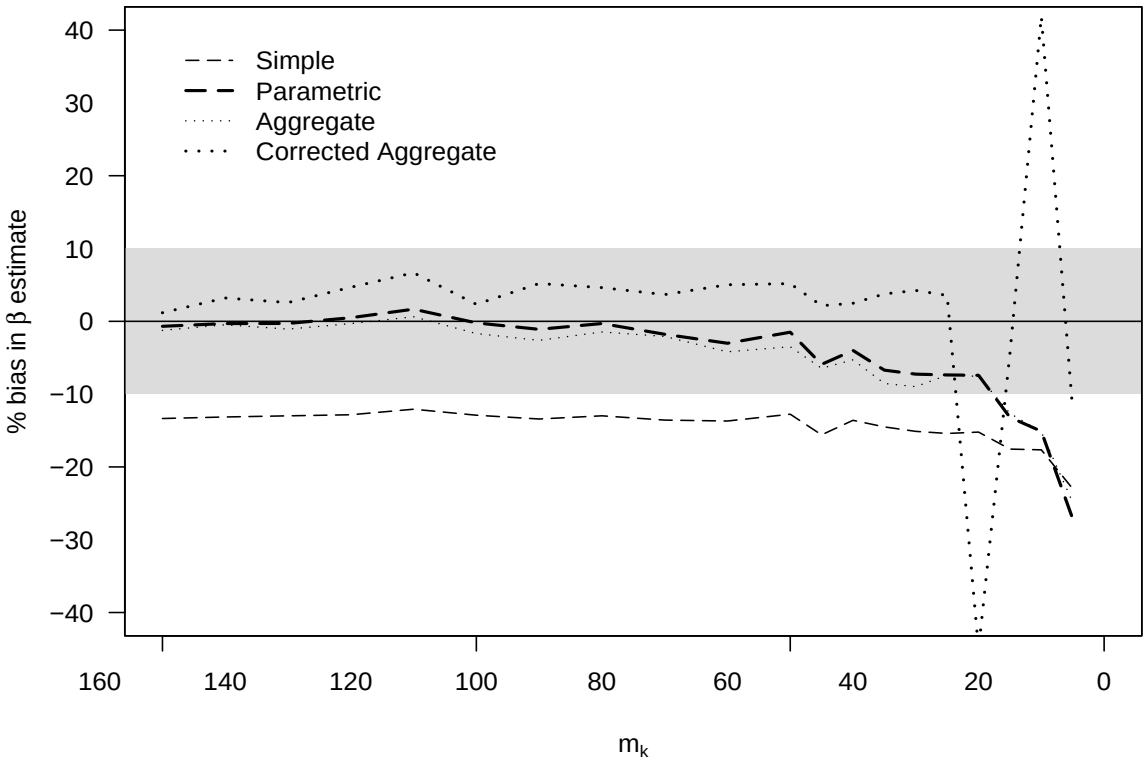

Figure 3: How the simple (light dashed), parametric (heavy dashed), aggregate (light dotted) and corrected aggregate (heavy dotted) models perform as a function of sample size m_k in each area. The within-area distribution is a normal distribution with variances increasing with the mean; further details appear in the main text.

where DP is a Dirichlet Process, F_{0k} is a known baseline distribution function and α_k is the strength of belief in F_{0k}. Both the unknown distribution F_k and the baseline distribution F_{0k} may include dependence between exposures.

We now wish to derive the implied aggregate disease model when we have a sample $\mathbf{X}_k^{m_k}$ of exposures from area k; that is, the model $Y_k|\mathbf{X}_k^{m_k}, \alpha_k, F_{0k}$. Following results given in Ferguson (1973), it can be shown that the mean of the disease counts is given by

$$E[Y_k|\mathbf{X}_k^{m_k}, \alpha_k, F_{0k}] = n_k\theta_k = n_k\left\{w_k\theta_{0k} + (1-w_k)\widehat{\theta}_{Ak}\right\}, \tag{4.2}$$

a weighted combination of the expectation of Y_{ki} under the prior distribution F_{0k}:

$$\theta_{0k} = E[\exp(\beta_0 + \beta_1 X_{ki})|F_{0k}], \tag{4.3}$$

and under the aggregate model, based on the sample data:

$$\widehat{\theta}_{Ak} = \frac{1}{m_k}\sum_{j=1}^{m_k}\exp(\beta_0 + \beta_1 X_{kj}), \tag{4.4}$$

with weights given by

$$w_k = \frac{\alpha_k}{\alpha_k + m_k}. \tag{4.5}$$

This model can be fitted via an estimating equations approach in the same way as the aggregate model (Prentice and Sheppard, 1995). While it is possible to derive an expression for the variance if we assume prior independence for the exposures, this is complex and it is more practical to use a working variance matrix (for example, assuming a constant variance) and sandwich estimation for inference (as in Prentice and Sheppard, 1995).

Model (4.2) is appealing as it represents a compromise between the parametric and aggregate models. When sample sizes are small the sample data are combined with a distributional assumption, borrowing strength from the prior and stabilising the estimator. When samples are larger they give more accurate information on the within-area distribution and will adjust inadequacies in the prior information, such as incorrectly-specified moments.

4.2 Adjustment for Small Samples

The above model will work well when we have moderate sample sizes and fairly accurate prior information. However, when the samples are small it combines the prior data, which may not be directly applicable to the current data, with the covariate data, which we have seen produces biased estimates when samples are small. Unless the prior information is very accurate and the α_k are chosen to be correspondingly large (in which case there is little need for the sample of individual data), estimates will be biased due to finite sampling bias in the estimating equations.

Following the corrected aggregate method, we may use a corrected version of the hybrid model to adjust for the finite sampling bias. We use an adjusted estimating equation similar to the corrected aggregate model which requires estimation of extra terms; we estimate these terms with a combination of both the data and the prior distribution. Further details are given in the Appendix. The corrected aggregate model provides a consistent estimator of β_1 with asymptotic normality as $m_k \to n_k$. Since the hybrid model tends to the aggregate model as $m_k \to n_k; n_k \to \infty$ we also have a consistent, asymptotically normal estimator for the hybrid model (provided the prior has the correct support).

This corrected version has three benefits. Firstly, the hybrid model is generally less biased for smaller samples than the uncorrected aggregate approach. Secondly, when the samples are small, we combine the prior data with the less biased estimation of the corrected aggregate method. Finally, estimation of the adjustment term which is responsible for the instability in the corrected aggregate approach is made less variable by smoothing it with the prior data.

4.3 Choice of Prior Distribution

The prior distribution function F_{0k} may be considered analogous to specifying a parametric within-area exposure distribution, and the weights (4.5) shows that the precision parameter

α_k may be viewed as the sample size associated with the specification F_{0k}. The distribution F_{0k} may be obtained from historical data, for example previous census data for demographic exposures, or previous years' pollution measurements for environmental exposures.

Obtaining suitable prior information may not be straightforward. In many cases, if we have reasonably-sized samples which will dominate, the prior distributions could be chosen to be independent normal distributions as a reasonable approximation. Previous simulations with the parametric approach (Wakefield and Salway, 2001) suggest that for mildly skewed distributions assuming a normal distribution will often give reasonable results, particularly when the exposure effect is small. This suggests that when we have a reasonable amount of individual data, specification of a suitable prior distribution for the hybrid approach should concentrate on obtaining good quality prior information about the first two moments of the within-area distributions.

In situations where sample data are sparse, estimation based on the hybrid approach will be sensitive to the choice of prior distribution and the prior moments. While prior information on exposure means may often be readily available, it may be difficult to obtain good prior values for within-area variances. In Section 3.2, we saw the importance of the within-area variances and in particular we require a characterisation of the within-area mean-variance relationship. This suggests that in specifying F_{0k} it is the relationship between the means and variances that is important, rather than the variances themselves. Where prior data are available, it may be beneficial to use smoothed versions of the variances, based on the mean-variance relationship. If variances are not available, it may sometimes be possible to specify prior data in terms of within-area means and a functional form for the relationship between the means and variances. At the very least this provides scope for a sensitivity analysis.

5 Simulations

5.1 Simulation Framework

In this section we describe the results of a simulation study, looking at a range of scenarios. The base scenario assumes a total population of $n_k = 2000$ in each area, for $K = 50$ areas. Individual data X_{ki}, exposure, and Y_{ki}, mortality, were generated within each area according to the model

$$\begin{aligned} \mu_k &\sim U(10, 15) \\ \sigma_k^2 &= a + b\mu_k \\ X_{ki} &\sim_{\text{ind}} N(\mu_k, \sigma_k^2) \\ Y_{ki} &\sim_{\text{ind}} \text{Bern}(p_{ki}) \\ \log(p_{ki}) &= \beta_0 + \beta_1 X_{ki} \end{aligned}$$

with $a = -3$ and $b = 0.4$, so within-area variances increase with the means, and $1 < \sigma_k^2 < 3$. This gives a between to within-area variability ratio in exposure of 1.1. The risk parameters were chosen as $\beta_0 = 5$, $\beta_1 = -\log(2)$ giving a negative relationship between exposure and mortality (β_0 was chosen so that most individuals would have $p_{ki} < 0.1$, that is, a rare disease).

The generated individual data were then aggregated to the area level to give disease counts Y_k in each area. Samples of X_{ki} of size $m_k = 20$ and $m_k = 5$ were taken in each area. For the parametric ecological model, the sample means and variances X_k and s_k^2 were calculated from these samples.

In addition to this base scenario, we considered three other scenarios as listed in Table 1 which have different choices for the between to within-area variability ratio, ranging from as large as 2, to the smallest of 0.06 (which is similar to the ratio observed in the income data). Previous

Table 1: The four different scenarios used to generate simulation data, using $\beta_1 = \log(2)$.

No.	Distribution of X_{ki}	Distribution of μ_k	Mean-Variance Relationship	Range of σ_k^2	Between:Within Ratio
1	Normal	U(10,15)	$\sigma_k^2 = -3 + 0.4\mu_k$	(1,3)	1.1
2	Normal	U(10,15)	$\sigma_k^2 = -3.7 + 0.38\mu_k$	(0.1,2)	2.1
3	Normal	U(11,15)	$\sigma_k^2 = -10 + \mu_k$	(1,5)	0.47
4	Normal	U(13.5,14.5)	$\sigma_k^2 = -12.5 + \mu_k$	(1,2)	0.06

authors (for example Richardson et al., 1987; Prentice and Sheppard, 1995) have suggested that ecological analyses work best when this ratio is large in order to exploit between-area mean contrasts.

To each dataset we fitted a range of different models; a full list appears in Table 2. Firstly we fit the simple ecological, the parametric normal, the aggregate and the corrected aggregate models. The simple and parametric models allow for overdispersion, via a quasi-likelihood method.

For the hybrid models, we used a normal prior distribution for F_{0k} and considered four different possibilities for prior moments. First, we used the true within-area moments as the prior values. Although unrealistic in practice, this choice represents the best case scenario for the hybrid model. Second, we used the true moments with random normal variation added. This represents the situation where we have good, unbiased information about both parameters from another source. Third, we repeated this with larger variation; this represents the situation where we have poor unbiased information of both parameters. Lastly, we used the good priors from the second situation and smoothed the variance, as a function of the mean, using a loess smoother. In all cases, we used the corrected hybrid model.

Table 2: Models fitted to each of the simulated datasets. For each hybrid model the prior distribution is $F_{0,k} \sim N(\mu_{0k}, \sigma^2_{0k})$ with μ_{0k} and σ^2_{0k} as given in the table.

Model	Description
Simple	Basic model (3.1), using sample means.
Parametric	Parametric model (3.2), with normal exposures, and sample means and variances
Aggregate	Aggregate model using samples of data
Aggregate (C)	Aggregate models using samples of data, corrected for small samples
$\text{Hybrid}_{\text{true}}$	True moments as prior information: $\mu_{0k} = \mu_{xk}$ and $\sigma^2_{0k} = \sigma^2_{xk}$
$\text{Hybrid}_{\text{good}}$	Good estimates of the true moments as prior information: $\mu_{0k} = \mu_{xk} + \epsilon_k, \quad \sigma^2_{0k} = \sigma^2_{xk} + \delta_k, \quad \epsilon_k, \delta_k \sim N(0, 0.25^2)$
$\text{Hybrid}_{\text{poor}}$	Poor estimates of the true moments as prior information: $\mu_{0k} = \mu_{xk} + \epsilon_k, \quad \sigma^2_{0k} = \sigma^2_{xk} + \delta_k, \quad \epsilon_k, \delta_k \sim N(0, 0.5^2)$
$\text{Hybrid}^{(s)}_{\text{good}}$	As for $\text{Hybrid}_{\text{good}}$, smoothing σ^2_{0k} using a loess smoother

5.2 Baseline Scenario

Table 3 reports results for 100 simulations for the baseline scenario. We see that as expected the simple, parametric and uncorrected aggregate models all have substantial bias, between around 15–20%. A reduced summary for the other scenarios is given in Table 4. Throughout this section results in the text are given for a sample size of $m_k = 20$, with corresponding results for $m_k = 5$ in brackets.

For these simulations we expect within-area variability bias to be negative and biased towards the null. When moments are estimated from small samples we see increased bias towards the null for the simple, parametric and uncorrected aggregate models. Plummer and Clayton (1996) speculated that the measurement error introduced by estimating moments would be such that a parametric ecological model would be worse than the simple ecological model, in a mean-squared error sense, but for the simulations here this is not the case. For the corrected aggregate method the algorithm either fails to converge or converges to the wrong value in 18% of simulations with $m_k = 20$ (for $m_k = 5$, 46%). Even excluding these, the mean-squared error is larger than for all other methods, because of the inflated variance of the estimator.

In general the hybrid model performs well, with similar-sized bias to the corrected aggregate method, but with smaller standard errors. Both the bias and the mean squared error is reduced compared to previous models. In addition, the convergence of these models is much improved over the corrected aggregate model; over 95% convergence for the hybrid compared to 82% (for $m_k = 5$, 85% compared to 54%). The aggregate model converged to the wrong value in 2 simulations (for $m_k = 5$, 9 simulations); this never occurred with the hybrid model.

Table 3: Results for simulation scenario 1. The corrected aggregate method results exclude 2 and 9 simulations for $m_k = 120$ and $m_k = 5$ respectively, where the algorithm converged to the wrong value. Including these gives mean biases of -14% and -32%. The true value of β_1 is -0.69 ($e^{\beta_1} = 0.5$) and the nominal confidence interval coverage is 95%.

Model	β_1	s.e.(β_1)	exp(β_1)	CI coverage %	% bias	MSE $\times 10^3$	% Convergence
Subsamples size $m_k = 20$							
Simple	-0.56	0.025	0.57	0	-20	20.3	100
Parametric	-0.59	0.034	0.55	19	-14	12.2	100
Aggregate	-0.59	0.033	0.56	11	-15	13.4	100
Aggregate (C)	-0.69	0.096	0.50	83	0	28.4	82
$\text{Hybrid}_{\text{true}}$	-0.68	0.054	0.51	84	-1	8.1	97
$\text{Hybrid}_{\text{good}}$	-0.69	0.059	0.50	90	0	8.9	97
$\text{Hybrid}^{(s)}_{\text{good}}$	-0.70	0.059	0.50	88	1	9.3	99
$\text{Hybrid}_{\text{poor}}$	-0.63	0.062	0.53	87	-9	10.8	99
Subsamples size $m_k = 5$							
Simple	-0.50	0.036	0.61	0	-28	39.1	97
Parametric	-0.47	0.040	0.63	0	-32	53.7	100
Aggregate	-0.49	0.038	0.61	0	-29	44.6	100
Aggregate (C)	-0.63	0.141	0.54	87	-9	46.5	54
$\text{Hybrid}_{\text{true}}$	-0.68	0.055	0.51	86	-2	7.4	83
$\text{Hybrid}_{\text{good}}$	-0.61	0.046	0.54	54	-12	11.1	100
$\text{Hybrid}^{(s)}_{\text{good}}$	-0.62	0.045	0.54	54	-11	10.2	100
$\text{Hybrid}_{\text{poor}}$	-0.64	0.081	0.53	94	-8	13.4	85

Table 4: Summary of results for other scenarios on hybrid models. Results given are average estimate of β_1, (average standard error) and average % bias in bold.

Scenario	Variability ratio	$\text{Hybrid}_{\text{true}}$			$\text{Hybrid}_{\text{good}}$			$\text{Hybrid}_{\text{poor}}$		
Subsamples size $m_k = 20$										
2	2.1	-0.68	(0.035)	**-2%**	-0.67	(0.036)	**-3%**	-0.63	(0.047)	**-9%**
3	0.5	-0.58	(0.126)	**-16%**	-0.59	(0.133)	**-14%**	-0.56	(0.118)	**-20%**
4	0.06	-0.36	(0.243)	**-48%**	-0.36	(0.222)	**-48%**	-0.26	(0.153)	**-63%**
Subsamples size $m_k = 5$										
2	2.1	-0.68	(0.034)	**-2%**	-0.67	(0.039)	**-4%**	-0.52	(0.058)	**-24%**
3	0.5	-0.56	(0.121)	**-20%**	-0.55	(0.110)	**-21%**	-0.48	(0.073)	**-31%**
4	0.06	-0.35	(0.170)	**-49%**	-0.32	(0.138)	**-53%**	-0.16	(0.098)	**-76%**

5.3 Between to Within-Area Variation Ratio

The extent of the bias depends most strongly on the between to within-area variance ratio; as this decreases the bias increases (noted also by Richardson et al., 1987). Figure 4 illustrates the extent of the bias as the between to within-area variability increases and Table 4 contains numerical summaries. This is the case for both the parametric and aggregate methods, and the new hybrid model. None perform well when this ratio is less than 1, and these simulations suggest that attempting ecological inference in such cases is not advisable. In the extreme case of scenario 4, with a between to within-area variability ratio of 0.05, even using the true moments as the prior information still results in around 50% bias. As the variability ratio decreases these model exhibit problems in terms of convergence, with convergence rates of around only 50% for a variability ratio of 0.05.

When the ratio is around 1 (scenario 1) the bias is reduced to within 2% if we have both moderate-sized samples and good quality prior information. With only one of these, the bias is

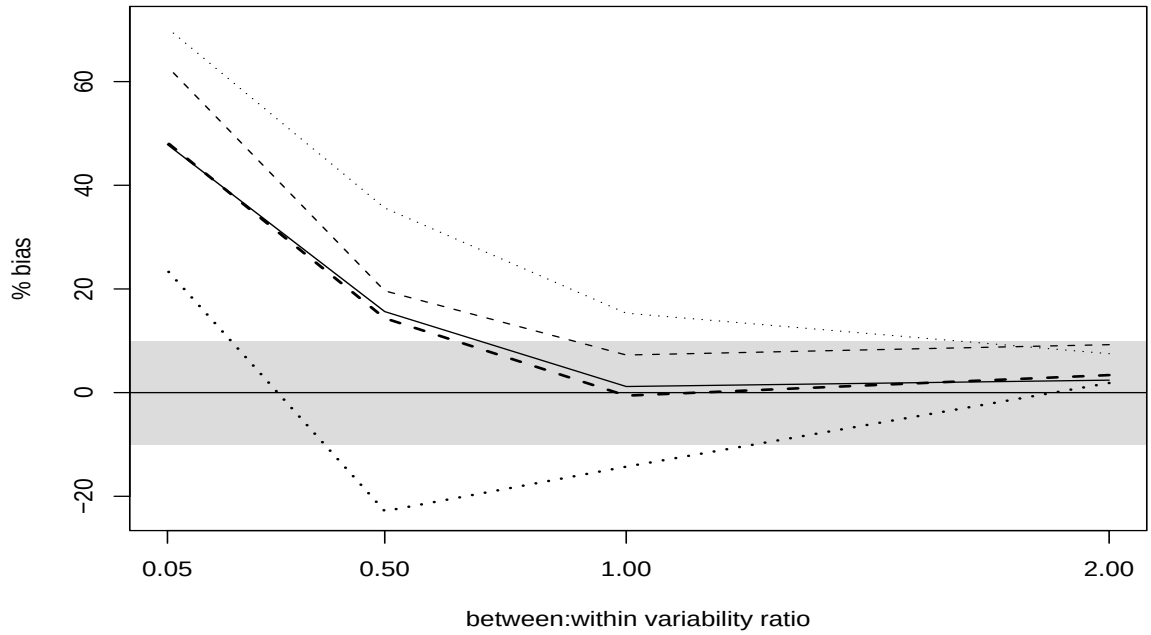

Figure 4: Comparison of % bias for between to within area variance ratios for the following models: aggregate (light dotted), corrected aggregate (heavy dotted), $\text{Hybrid}_{\text{true}}$ (solid), $\text{Hybrid}^{(s)}_{\text{good}}$ (heavy dashed) and $\text{Hybrid}_{\text{poor}}$ (dashed). The region of $\pm 10\%$ bias is shaded.

higher at around 10%. Ideally the ratio should be 2 or higher; that is, more variability between areas than within. In this case, even if samples are small we can remove nearly all the bias provided we have good quality prior information.

5.4 Quality of Prior Information

If the prior information is 'perfect' and takes the form of the true moments (that is, the values μ_k, σ_k^2 from which the data were generated) we obtain extremely good estimates of β_1, with bias of 1–2% for both sample sizes. For $m_k = 20$ using good unbiased estimates ($\text{Hybrid}_{\text{good}}$) generally gives very similar results. For smaller samples, the bias is increased and the model is

only feasible when the between to within-area variability ratio is large.

The hybrid model essentially pulls the sample data towards the prior moments; as the quality of the prior becomes worse, either in terms of how well they are estimated, or in terms of bias in the prior moments, estimates of β_1 become more biased. To some extent the quality of the prior data may be reflected in the parameter α_k, but a small value of α_k will effectively result in using the corrected aggregate method with its instability problems for small samples. Since the method depends so strongly on the prior information, when these data are biased (for example if the prior means are consistently over-estimated) we would expect estimation to be poorer; in additional simulations, not shown, this was indeed the case.

5.5 Sample Size

Sample size affects the performance of the models. Larger samples will partially compensate for poor quality prior information; for example $\text{Hybrid}_{\text{poor}}$ is less biased for $m_k = 20$ than for $m_k = 5$. As sample sizes become much larger ($m_k > 100$) there is nearly no bias, unless the within-area distribution is extremely skewed. We conclude that in the situation of interest, when we only have very small samples, we require good quality prior and a large variability ratio for an ecological study to be feasible. Figure 5 illustrates how the bias depends on the sample size; the aggregate models from Figure 3 are shown for comparison.

5.6 Other Factors

Smoothing the prior variances to focus on the underlying mean-variance relationship has only a very small effect, since it often picks up erroneous patterns. The results are inconclusive, but slight improvements suggest that in general smoothing is preferable, especially if the relationship between the means and variances is very strong. We also looked at additional simulations which

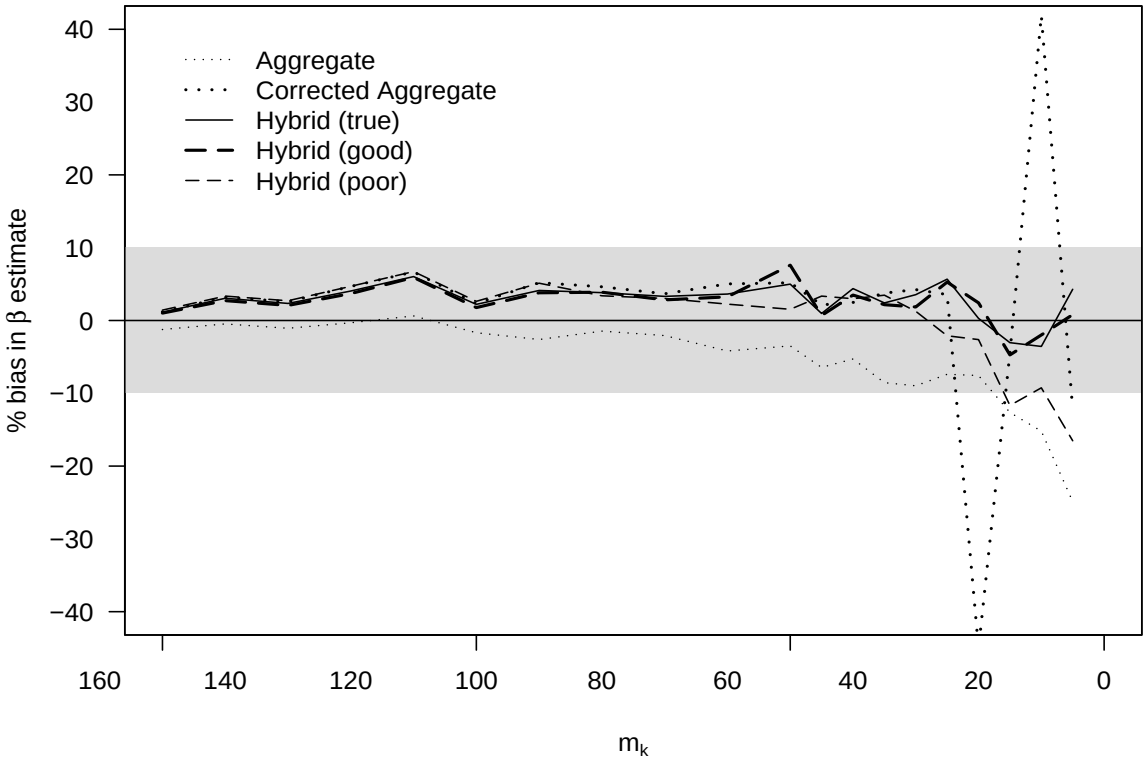

Figure 5: The bias in the hybrid models as a function of the sample size: $\text{Hybrid}_{\text{true}}$ (solid); $\text{Hybrid}^{(s)}_{\text{good}}$(heavy dashed) and $\text{Hybrid}_{\text{poor}}$ (dashed). For comparison, Aggregate (dotted) and adjusted aggregate (heavy dotted).

investigated the mean-variance relationship. Factors such as additional variability, or a loglinear (rather than linear) mean-variance relationship made very little difference to the results. Finally, these simulations show the effects of bias on a negative effect parameter; if $\beta_1 > 0$ similar patterns emerge, although the bias is more complex in this situation as positive and negative biases from different sources may to some extent cancel each other out.

5.7 Well-estimated mean

One interesting scenario is when well-estimated area means are available from some source (for example, routinely collected data such as from the census), and the purpose of the additional samples is to estimate the within-area variability. This is a potentially useful situation where we might expect to further reduce ecological bias.

One way to incorporate such data into the hybrid model is to use the good estimate of the mean as the prior mean. Depending on how the well-estimated mean has been derived, there may be a small possibility that an individual appears both in the prior and in the sample. However, with very small samples and reasonably large areas the probability of this is extremely small and may be often be considered negligible.

We investigated the performance of the hybrid model for this situation in scenario 1. We generated a separate sample of 100 from the within-area distributions and used the mean of this sample as the prior mean. Good and poor prior information on within-area variances was simulated as before; the results are shown in Table 5.

Good prior information about the mean only is good enough to reduce bias to under 3%, even when the samples are small and the prior information about the variance is poor. These results are consistent across the other scenarios, with increased bias for a smaller variability ratio, suggesting that the hybrid approach is beneficial in this scenario.

Table 5: Results for scenario 1, assuming that the within-area mean is well-estimated.

Model	β_1	s.e.(β_1)	exp(β_1)	CI coverage %	% bias	MSE $\times 10^3$
TRUTH	-0.69	-	0.5	95	0	0
Subsamples size $m_k = 20$						
Simple	-0.58	0.020	0.56	0	17	14.7
Parametric	-0.62	0.029	0.54	35	10	6.4
$\text{Hybrid}_{\text{good}}$	-0.68	0.054	0.51	88	1	8.0
$\text{Hybrid}_{\text{poor}}$	-0.68	0.051	0.51	80	3	8.0
Subsamples size $m_k = 5$						
Simple	-0.58	0.020	0.56	0	17	14.3
Parametric	-0.54	0.029	0.58	0	22	24.4
$\text{Hybrid}_{\text{good}}$	-0.67	0.053	0.51	87	3	7.3
$\text{Hybrid}_{\text{poor}}$	-0.68	0.061	0.51	93	2	7.3

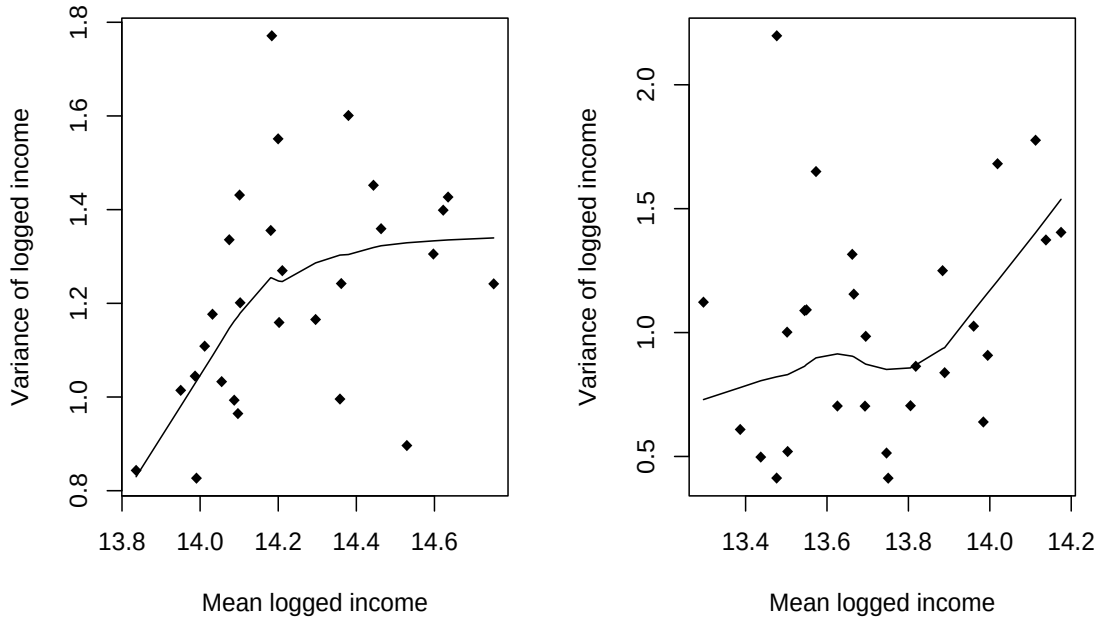

Figure 6: Relationship between within-area means and variances, smoother imposed.

6 Income data

In this section we compare the hybrid model to previous models on the data described in Section 2. The relationship between the within-area means and variances is plotted in Figure 6; in both cases, the variance increases with the mean.

The between to within area variability ratio is very small in this data set (around 0.05), and the simulations in Section 5 suggest that an ecological analysis of this data is inappropriate. For illustrative purposes we will fit the models, anticipating that we may experience problems with some of the methods, and not all the bias will be removed. Of our two datasets, we expect better estimates of exposure effect for all men than for men over 64, since the sample sizes are larger.

Table 6: Comparison of models for income data.

	All men			Men over 64		
Model	β_1	95% CI	exp(β_1)	β_1	95% CI	exp(β_1)
Simple	-0.305	(-0.56 , -0.05)	0.74	-0.061	(-0.14, 0.01)	0.94
Parametric	-0.266	(-0.53 , 0.00)	0.77	-0.058	(-0.14, 0.02)	0.94
Aggregate	-0.295	(-0.54 , -0.05)	0.75	-0.062	(-0.14, 0.02)	0.94
Aggregate(C)	-0.435	(-0.85 , -0.02)	0.65	1.321	(-0.20, 2.84)	3.75
Hybrid	-0.318	(-0.70, 0.06)	0.73	-0.826	(-1.45, -0.21)	0.44

We fit the simple ecological, the parametric normal, the aggregate and the corrected aggregate models. For the hybrid model, we specify a normal prior distribution F_{0k} with prior mean and variance (on the log base 2 scale). Areas were ranked according to socioeconomic status, and information on mean income from other sources (the Family Resources Survey 2002/3 (Department for Work and Pensions, 2002a), and The Pensioners' Income Series 2002/3 (Department for Work and Pensions, 2002b)) was combined with prior beliefs about the mean-variance structure to give prior data at the regional level. We use a prior sample size of $\alpha_k = 10$, with corresponding weights on the prior of w_k between 0.04 and 0.14 for all data, and between 0.2 and 0.8 for the over 64s data.

Table 6 summarises the results of the different models. For the first dataset comprising all men, we would expect all models to be biased to some extent, because of the small between to within-area variability ratio. However, since the samples are moderately large (m_k=60–216) we would expect only small differences between the parametric, aggregate and hybrid models. This is the case, with slightly larger standard errors for the hybrid model. All relative risk estimates suggest a 65–75% reduction in risk for every doubling of income, although this reduction is not significant for the hybrid model.

For the over 64 data the smaller sample sizes (m_k=3–44) cause differences between the models. The parametric and aggregate models suggest that there is no significant effect of income on health. The corrected aggregate model estimates a positive association between income and health; this seems implausible, which suggests this is a practical example of convergence to the wrong value. The hybrid model suggests that the true relative risk is closer to 0.4, representing a reduction in mortality risk of 44% for every doubling of income; substantially lower than the other models would suggest. This difference is also found to be significant, despite the much larger standard error.

7 Discussion

The key to using ecological data for individual inference is the availability of individual data. In this paper we have presented a new model that combines ecological and individual data, and reduces within-area variability bias in situations when individual data are difficult to obtain, and only small samples are available.

The hybrid model can be seen as a compromise between the existing parametric and corrected aggregate models which perform poorly for small sample sizes; the former produces biased estimates, and the latter is unreliable and unstable. With suitable prior information the hybrid model can reduce within-area variability bias to within 5%. A key factor is the between to within-area variability ratio which ideally should be 2 or higher; when this ratio is small (less than 1) ecological analysis is inadvisable. If it is known in advance that the ratio is likely to be small, we can compensate to some extent by increasing the sample sizes and collecting more data.

The sample size required for accurate estimation with the hybrid model depends not only on the exposure variability ratio but also on the quality of the prior information and the within-area

distribution. If resources are limited we should choose larger samples for those areas where the prior information is known to be poor, those with larger within-area variability and those where the exposure distribution is likely to be heavily skewed. In contrast, a nearly homogenous area with good prior information requires little individual data. Figure 5 suggests that a good strategy is to use the aggregate model for large samples ($m_k > 100$), the corrected aggregate or uncorrected hybrid models for moderate samples ($50 < m_k < 100$), and the corrected hybrid model for small samples ($50 < m_k < 10$). Samples of less than 10 will rarely be feasible, unless within-area variability is small. The hybrid model can be adjusted for these different strategies, on an area-by-area basis, by altering the prior sample size α_k; setting $\alpha_k = 0$ in some areas, for example, will have the effect of discarding prior information and using the sample data exclusively for those areas. Whatever strategy is employed, some sensitivity analysis to the choice of α_k is a good idea.

The hybrid model relies on the presence of good quality prior information about the within-area distributions. The most important aspects of the prior are to have fairly accurate information on the area means, and to capture accurately the relationship between the mean and the variance. For very small samples, the model may exhibit problems if the within-area distribution is mis-specified. Additional simulations not presented here explored the use of a normal prior assumption when the true distribution was lognormal; the results showed extreme instability in the algorithm. This is due primarily to the correction for finite sampling bias, where the extra term required is now estimated from a combination of the highly variable data and the incorrect prior distribution. This suggests that when little is known about the true form of the prior distribution, larger samples will be required to be able to use the uncorrected hybrid model.

We have not discussed how this model may be extended to deal with multiple exposures and confounders. While it is straightforward to write down a suitable model, in practice fitting this model will be more complex and potentially problematic. The individual data will need to char-

acterise the within-area joint exposure-confounder distribution, capturing the mean-variance-covariance structure for all covariates. One consequence is that individual covariate data will be required on the same set of individuals; for example, combining data from several surveys will not be possible without assuming independence between covariates. Further research is required to investigate the use of the hybrid model in these more complex situations. However, the hybrid model is ideally suited to the semi-ecological study, where individual data are available on the outcome and confounders, with exposure information coming from ecological data.

Using ecological data for individual inference is problematic. The hybrid model is designed to reduce ecological bias in a situation where otherwise no reliable analysis is possible; it should never be used as a substitute for collecting large samples of individual data where this is possible, or in place of a well-designed individual study.

A Appendix

We give details of the hybrid model corrected for finite sampling bias, assuming constant variance of Y_k. The approach follows closely that of Prentice and Sheppard (1995) for the corrected aggregate method.

The expectation of the estimating equation for the uncorrected hybrid model is given by

$$E\left[\sum_{k=1}^{K} D_k^T(y_k - n_k\widetilde{\theta}_k)\right] = -\sum_{k=1}^{K}(1-w_k)^2\frac{(n_k-m_k)}{m_k(n_k-1)}(n_k S_k^T - D_k^T\theta_k), \qquad \text{(A.1)}$$

where θ_k is the average risk and $\widetilde{\theta}_k = w_k\theta_{0k} + (1-w_k)\widehat{\theta}_{Ak}$ is the estimate under the hybrid model given by (4.2), D_k is the $p \times 1$ vector of derivatives, $D_k = \frac{\partial}{\partial\boldsymbol{\beta}}\theta_k$, and $S_k = D^T\theta_k$. The summation term in this expression is the same as in the aggregate estimating equation, but is given less weight, since there is finite sampling bias only in the data part of the model, and not from the prior information. So we expect the hybrid approach to generally suffer less finite

sampling bias than the aggregate for similar sized samples (overall, the bias depends on the accuracy of the prior).

The correction factor is based on an estimated version of (A.1). However, unlike the corrected aggregate approach where the estimation of this term is highly variable, we can now estimate these terms from a combination of both the data *and* the prior distribution:

$$\begin{aligned} \widetilde{D}_k &= w_k D_{0k} + (1 - w_k)\widehat{D}_{Ak} \\ \widetilde{S}_k &= w_k S_{0k} + (1 - w_k)\widehat{S}_{Ak} \end{aligned} \tag{A.2}$$

where D_{0k} and S_{0k} are the expected values of D_k and S_k under the prior distribution F_{0k}, and D_{Ak} and S_{Ak} are the expected values under the aggregate model using $\mathbf{X}_k^{m_k}$:

$$\begin{aligned} \widehat{D}_{Ak} &= \frac{1}{m_k}\sum_{i=1}^{m_k} \begin{pmatrix} 1 \\ X_{ki} \end{pmatrix} e^{\boldsymbol{\beta}^T X_{ki}} \\ \widehat{S}_{Ak} &= \frac{1}{m_k}\sum_{i=1}^{m_k} \begin{pmatrix} 1 \\ X_{ki} \end{pmatrix} e^{2\boldsymbol{\beta}^T X_{ki}} \end{aligned} \tag{A.3}$$

For example, for a normal prior distribution $X_{ki}|F_{0k} \sim N(\mu_{0k}, \sigma^2_{0k})$ we have

$$\begin{aligned}
\theta_{0k} &= E[\exp(\beta_0 + \beta_1 X_{ki})|F_{0k}] \\
&= \exp(\beta_0 + \beta_1\mu_{0k} + \beta_1^2\sigma^2_{0k}/2) \\
D_{0k} &= E\left[\begin{pmatrix} 1 \\ X_{ki} \end{pmatrix} \exp(\beta_0 + \beta_1 X_{ki})|F_{0k}\right] \\
&= \begin{pmatrix} 1 \\ \mu_{0k} + \beta_1\sigma^2_{0k} \end{pmatrix} \exp(\beta_0 + \beta_1\mu_{0k} + \beta_1^2\sigma^2_{0k}/2) \\
S_{0k} &= E\left[\begin{pmatrix} 1 \\ X_{ki} \end{pmatrix} \exp(2\beta_0 + 2\beta_1 X_{ki})|F_{0k}\right] \\
&= \begin{pmatrix} 1 \\ \mu_{0k} + 2\beta_1\sigma^2_{0k} \end{pmatrix} \exp(2\beta_0 + 2\beta_1\mu_{0k} + 2\beta_1^2\sigma^2_{0k}) \qquad \text{(A.4)}
\end{aligned}$$

So the corrected hybrid model is the solution to the corrected estimating equation,

$$\sum_{k=1}^{K} \widetilde{D}^T_{Ak}(y_k - n_k\widetilde{\theta}_k) + \widetilde{m}_k^{-1}(n_k\widetilde{S}^T_{Ak} - \widetilde{D}^T_{Ak}\widetilde{\theta}_{Ak}), \qquad \text{(A.5)}$$

with

$$\widetilde{m}_k = \frac{m_k(n_k - 1) - (1 - w_k)^2(n_k - m_k)}{(1 - w_k)^2(n_k - m_k)}, \qquad \text{(A.6)}$$

which is unbiased in expectation over all possible choices of subsample. This form may be used with a constant working variance matrix and sandwich estimation for empirical standard errors.

References

Department for Work and Pensions (2002a). *The Family Resources Survey 2002/03* (online edition), Department for Work and Pensions. Available from: `http://www.dwp.gov.uk/asd/frs/index/publications.asp`.

Department for Work and Pensions (2002b). *The Pensioners Incomes Series 2002/03* (online edition), Department for Work and Pensions. Available from: `http://www.dwp.gov.uk/asd/asd6/pensioners_income.asp`.

Ferguson, T. S. (1973). Bayesian analysis of some non-parametric problems. *Annals of Statistics*, 1:209–30.

Gravelle, H., Wildman, J., and Sutton, M. (2002). Income, income inequality and health: What can we learn from aggregate data? *Social Science and Medicine*, 54:577–89.

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*, 18:269–74.

Greenland, S. and Robins, J. (1994). Invited commentary: Ecologic studies - biases, misconceptions and counterexamples. *American Journal of Epidemiology*, 139:747–64.

Judge, K., Mulligan, J., and Benzeval, M. (1998). Income inequality and population health. *Social Science and Medicine*, 46:567–79.

McClements, L. (1977). Equivalence scales for children. *Journal of Public Economics*, 8:191–210.

Ott, W. R. (1994). *Environmental Statistics and Data Analysis*. CRC Press.

Plummer, M. and Clayton, D. (1996). Estimation of population exposure. *Journal of the Royal Statistical Society, Series B*, 58:113–26.

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 Elliott, P., Wakefield, J. C., Best, N. G., and Briggs, D. J., editors, *Spatial Epidemiology: Methods and Application*, chapter 11. Oxford University Press, Oxford.

Richardson, S., Stucker, I., and Hémon, D. (1987). Comparison of relative risks obtained in ecological and individual studies: Some methodological considerations. *International Journal of Epidemiology*, 16:111–20.

Salway, R. and Wakefield, J. (2004). A comparison of approaches to ecological inference in epidemiology, political science and sociology. In King, G., Rosen, O., and Tanner, M., editors, *Ecological Inference: New Methodological Strategies*, chapter 14. Cambridge University Press, New York.

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*, 15:1849–58.

UK Data Archive: SN4817 (2002). Office for National Statistics, *Vital Statistics for England and Wales, 2002* [computer file]. Colchester, Essex: UK Data Archive [distributor], February 2004. SN: 4817.

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

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

List of Tables

List of Figures