Ecological Inference for $2 \times 2$ Tables

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

Summary. A fundamental problem in many disciplines, including political science, sociology and epidemiology, is
the examination of the association between two binary variables across a series of $2 \times 2$ tables, when only the mar-
gins are observed, and one of the margins is fixed. There are two unobserved fractions that are of interest, with
only a single response per table, and it is this non-identifiability that is the fundamental difficulty lying at the heart
of ecological inference. Many methods have been suggested for ecological inference, often without a probabilistic
model; in this paper we clarify the form of the sampling distribution and critique previous approaches within a for-
mal statistical framework thus allowing clarification and examination of the assumptions that are required under
all approaches. A particularly difficult problem is choosing between models with and without contextual effects.
Various Bayesian hierarchical modelling approaches are proposed to allow the formal inclusion of supplementary
data, and/or prior information, without which ecological inference is unreliable. Careful prior choice within such
models is required, however, since there may be considerable sensitivity to this choice, even when the assumed
model is correct and there are no contextual effects. This sensitivity is shown to be a function of the number of
areas and the distribution of the proportions in the fixed margin across areas. By explicitly providing a likelihood
for each table, the combination of individual-level survey data and aggregate-level data is straightforward and we
illustrate that survey data can be highly informative, particularly if these data are from a survey of the minority
population within each area. This strategy is related to designs used in survey sampling and in epidemiology. An
approximation to the suggested likelihood is discussed, and various computational approaches are described. A
number of extensions are outlined including the consideration of multi-way tables, spatial dependence and area-
specific (contextual) variables. Voter registration/race data from 64 counties in the U.S. state of Louisiana are
used to illustrate the methods.

Keywords: Auxiliary Variables; Contextual Effects; Ecological Fallacy; Ecological Regression; Hierarchical Mod-
els; Identifiability; Markov Chain Monte Carlo; Method of Bounds; Missing Data; Neighbourhood Model; Extended
Hypergeometric Distribution; Simpson’s Paradox; Spatial Epidemiology; Survey Sampling.

1. Introduction

In this paper the problem of making individual-level inference from ecological data is considered. In particular
suppose we have a set of $2 \times 2$ tables in which the margins only are observed, for concreteness we suppose each
table corresponds to a different geographical area. This problem arises in many disciplines including political
science (Achen and Shively, 1995; King, 1997), sociology (Goodman 1953, 1959; Duncan and Davis, 1953)
and spatial epidemiology (Richardson and Montfort, 2000); King (1997) and Cleave et al. (1995) describe
further application areas. Although the shortcomings of ecological inference have been documented (e.g.,
Achen and Shively, 1995; Greenland and Robins, 1994; Freedman, 2001; Gelman et al., 2001), the continued
use of ecological data can be attributed to: the increased sample sizes and predictor ranges they provide;
their routine availability; their increased reliability when a sensitive question is asked; and the impossibility
of further data collection in historical contexts. The constituent tables may contain data on two distinct
binary variables recorded for each area, or the same binary variable recorded for each area at two time points
(for example, data from two election returns). For clarity we concentrate on the former case and label the
fixed binary predictor variable $x$, and the binary response variable $Y$.

The fundamental difficulty of ecological inference is that in the unaggregated data there are two quantities
of interest in each table (the number of $Y = 1$ responses at each of the two $x$ values), but only the total
number of responses is available in the aggregate data, and many sets of internal cells are consistent with
a given marginal table. As pointed out by Robinson (1950), association measures are inconsistent across
Jon Wakefield
different levels of aggregation. Selvin (1958) coined the term the ecological fallacy for the situation in which, “relationships between characteristics of individuals are wrongly inferred from data about groups”; the reversal of individual and aggregation conclusions is closely related to Simpson’s paradox, a connection we explore in Section 3. In geography, the inconsistency is known as the modifiable areal unit problem (MAUP), see for example Openshaw (1984). There is also a great deal of interest in the ecological fallacy in the epidemiology literature, particularly with respect to “geographical correlation studies”, see Piantadosi et al. (1988), Richardson et al. (1987), Greenland and Morgenstern (1989), Greenland (1992), Greenland and Robins (1994), Morgenstern (1998) and Wakefield (2003). In the social sciences great attention has been paid to contextual effects which exist when an individual response is influenced not only by individual characteristics, but also by characteristics of other individuals in a shared neighbourhood. Unfortunately, due to the aggregation, the estimation of such effects is extremely difficult and requires additional data, or is strongly dependent upon modelling assumptions.

The structure of this paper is as follows. In the next section we introduce a specific dataset on political registration/race in Louisiana that will be used to illustrate the ideas of the paper. In Section 3 we provide a framework for ecological inference and critically review a number of proposed approaches within this framework. In Section 4 a particular plausible likelihood is discussed, along with a normal approximation. Section 5 concentrates on hierarchical models and suggests a number of new modelling approaches. In Section 6 we discuss a number of the factors that must be considered when carrying out hierarchical modelling, using simulated data, and in Section 7 fit hierarchical models to the Louisiana data. Section 8 considers multiway tables and Section 9 describes how survey and aggregate data may be simply combined under the framework introduced in Section 4, and illustrates the benefits of such an approach. Section 10 contains a concluding discussion and an Appendix computational details.

2. Louisiana Data

To motivate our discussion we introduce a specific dataset for which \( Y = 0/1 \) represents the event Democrat/Republican registration, and \( z = 0/1 \) the event Black/White, representing race. The data were collected in the U.S. state of Louisiana in 1990 and we will examine these data at the level of the county, of which there are 64; these data are available at \( \text{http://gking.harvard.edu/data.shtml} \). For these data the internal cells are available, but in subsequent sections we will analyse the marginal data only, and assess the success of various approaches at providing accurate estimates of the internal cells.

We begin by examining the individual-level data. We let \( \hat{p}_{0i} \) and \( \hat{p}_{1i} \) denote the fractions of blacks and whites, respectively with Republican registration, and \( x_i \) the proportion black, in area \( i, i = 1, ..., 64 \). As is typical of applications such as this, the number of residents in each area is large: the median count is 16,613, the interquartile range 10,081–33,249, and the minimum and maximum counts are 4,421 and 217,967. Figure 1 shows the true fraction of white individuals with Republican registration plotted against the equivalent quantity for black individuals, by area. These points are superimposed on so-called tomography lines that give the admissible ranges of the fractions based on the margins alone; these will be described in greater detail in Section 5. Figures 2(a) and (b) give histograms of the fractions; the black fractions are much smaller and show much less variability across areas than their white counterparts. An important observation for ecological inference is that the majority of areas contain more whites than blacks. The fraction with Republican registration plotted against the proportion black, in each of the 64 counties. The fraction with Republican registration decreases as the proportion in the county who are black increases, with the obvious explanation being that
Fig. 1. Fractions of white populations with Republican registration, $\tilde{p}_{1i}$, plotted versus fractions of black populations with Republican registration, $\tilde{p}_{0i}$, for each of $i = 1, \ldots, 64$ counties, in Louisiana in 1990. Tomography lines, $\tilde{p}_{1i} = \tilde{q}_i/(1-x_i) - \tilde{p}_{0i}x_i/(1-x_i)$, defining the admissible ranges of these quantities from the marginal data only, are also displayed.

Fig. 2. For the Louisiana registration/race data: (a) Histogram of $\tilde{p}_{0i}$, (b) histogram of $\tilde{p}_{1i}$, (c) $\tilde{p}_{0i}$ versus $x_i$, (d) $\tilde{p}_{1i}$ versus $x_i$, where $\tilde{p}_{0i}$ and $\tilde{p}_{1i}$ denote the fraction of blacks and whites, respectively, with Republican registration, and $x_i$ denotes the proportion of blacks in county $i$, $i = 1, \ldots, 64$. Local smoothers have been added to (c) and (d).
blacks are less likely to be registered as Republican. Alternative explanations exist, however, in particular the same aggregate pattern could be observed if individual race had no effect, but individuals are registered as Republicans according to the race of those in their area, so that aggregate race is a contextual effect. The association could also be produced if whites are less likely to be registered as Republicans if resident in a county with more black residents, and/or if blacks are more likely to be registered as Republicans if resident in a predominantly white county. Hence, in the situation in which aggregate data only are available, a range of individual-level inferential explanations are consistent with the aggregate association, including the conclusion that blacks are more likely to have Republican registration than whites, in which case the original explanation is merely an example of the ecological fallacy.

3. Previous Approaches

3.1. The Sampling Distribution

In the social sciences, approaches to the ecological inference problem have often concentrated on imputing the missing cells in the constituent $2 \times 2$ tables. Hence particular approaches to the problem have been viewed (implicitly) in the context of a finite population, often without recourse to a formal statistical model, which causes difficulties in determining the properties of proposed estimation methods. Table 1 gives the notation for area $i$; in an ecological study, the marginal number registered Republican, $Y_i$, is observed, but $Y_{0i}$ and $Y_{1i}$, the numbers of blacks and whites with Republican registration, respectively, are unobserved. We assume that associated with each area/race classification there is a hypothetical infinite population with an associated probability of registration, and that the row totals in Table 1 are fixed. We let

$$ p_{0i} = \Pr(Y = 1|x = 0, i) \quad \text{and} \quad p_{1i} = \Pr(Y = 1|x = 1, i) $$

(1)
Table 1. Table summarising data in area $i$; in an ecological study the margins only are observed, $i = 1, ..., m$.

<table>
<thead>
<tr>
<th></th>
<th>$Y = 0$</th>
<th>$Y = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = 0$</td>
<td>$Y_{0i}$</td>
<td>$N_{0i}$</td>
</tr>
<tr>
<td>$x = 1$</td>
<td>$Y_{1i}$</td>
<td>$N_{1i}$</td>
</tr>
</tbody>
</table>

represent the probabilities of black and white Republican registration in area $i$, respectively, for $i = 1, ..., m$ areas (these correspond to the Republican fractions within the hypothetical infinite populations of blacks and whites). To clarify notation, we let $\tilde{p}_{0i} = y_{0i}/N_{0i}$ and $\tilde{p}_{1i} = y_{1i}/N_{1i}$ denote, retrospectively, the (unobserved) fractions of blacks and whites with Republican registration. We reserve $\hat{p}_{0i}$ and $\hat{p}_{1i}$ as estimates of the underlying probabilities/fractions. The distinction between the population probabilities $p_{0i}$ and $p_{1i}$ and the fractions $\tilde{p}_{0i}$ and $\tilde{p}_{1i}$ is important but has not always been made clear, leading to confusion.

If the internal cells were observed then, under the assumption of independent Bernoulli trials within each area and race category, the likelihood for the data of table $i$ consists of the product of a pair of independent binomial distributions:

$$Y_{ji}|p_{ji} \sim \text{Binomial}(N_{ji}, p_{ji}),$$

for $j = 0, 1$, $i = 1, ..., m$. To emphasize, for this derivation we have assumed that there is no clustering within areas due to unobserved variables; the latter would lead to overdispersion, due to the induced dependence (McCullagh and Nelder, 1989, p. 125). If there are contextual effects then the probabilities $p_{ji}$ will be a function of $x_i$.

The (marginal) probability of registration in area $i$ is given by

$$q_i = p_{0i}x_i + p_{1i}(1 - x_i),$$

where $x_i = N_{0i}/N_i$ and $1 - x_i = N_{1i}/N_i$ are, respectively, the (observed) proportions of blacks and whites in area $i$. The fundamental identifiability problem in ecological inference is that we wish to estimate $2m$ quantities from $m$ data points, the latter corresponding to the number of individuals who are registered in area $i$, $Y_i$, $i = 1, ..., m$.

At first sight it would appear that there is no information concerning $p_{0i}, p_{1i}$ in a table in which the margins only are observed. The consideration of the extreme cases reveals that this is not the case, however. If $N_{0i}$ ($N_{1i}$) is zero then we obtain individual-level information on $p_{1i}$ ($p_{0i}$) only. If $Y_i = N_i$ (0) then the fraction $\tilde{q}_i = 1 (0)$ which suggests that $p_{0i}$ and $p_{1i}$ are both close to one (zero). The strength of belief in the closeness to one of each probability is not in general symmetric, however. If, for example $N_{0i}$ is large in relation to $N_{1i}$ then the information that $p_{0i}$ is close to 1 (0) is stronger than the information that $p_{1i}$ is close to 1 (0). We also note that if $N_{0i} = N_{1i}$, then we have no information in the data alone to discriminate between $p_{0i}$ and $p_{1i}$. This provides a hint that the distribution of $x_i$ across areas is likely to be important since it determines the symmetry in the information available for the black and white probabilities.

Some insight into the difficulties of ecological inference may be gained by considering the nature of conditional inference for a single $2 \times 2$ table. A hypothesis test that the odds ratio is unity may be carried out using Fisher’s exact test, and estimation follows from a conditional likelihood that depends on the odds ratio only. At first sight, the conditional view would appear to suggest that nothing can be gained from ecological data with respect to estimation of the odds ratio, as the margins are conditioned upon since they contain “little information” about this parameter. The margins are not ancillary, however, and there is some information contained within them (see Plackett (1977) for further discussion), particularly when one has multiple tables, with an assumption of “similarity” across tables. Altham (1969) provides a Bayesian examination of Fisher’s exact test, and in particular shows that the test does not correspond to estimation with a uniform prior on the internal cell probabilities, but rather to a prior that favours association, albeit weakly.

Under the sampling scheme that leads to (2), and the assumption that the row totals $N_{0i}$ and $N_{1i}$ are fixed, the true likelihood in the case in which $(Y_{0i}, Y_{1i})$ are unobserved, is the convolution of binomial distributions.
The convolution likelihood given by (4) has rarely been explicitly considered in the ecological inference literature, though Achen and Shively (1995, p. 46) provide a good discussion, and McCullagh and Nelder (1989, p. 338) use this form in the situation of common registration probabilities across tables, and consider two approaches to estimation. In the first, a quasi-likelihood approach, estimation is carried out using the first two moments of \( Y_i \) given by (3). Under both this formulation and the convolution model we have \( \text{var}(\bar{y}) = \text{var}(\bar{y}) \). For the binomial model

\[
\text{var}(Y_i)_{p_{0i}, p_{1i}} = N_i q_i (1 - q_i),
\]

while for the convolution we have

\[
\text{var}(Y_i)_{p_{0i}, p_{1i}, N_i} = N_i q_i (1 - q_i) \left[ 1 - \frac{\tau_i^2}{\tau_i^2} \right],
\]

where \( 0 \leq \tau_i^2 = \frac{\text{var}(P_i)}{q_i (1 - q_i)} \leq 1 \), and \( P_i \) is a random variable taking the value \( p_{0i} \) with probability \( x_i \) and \( p_{1i} \) with probability \( 1 - x_i \). Hence

\[
\text{var}(Y_i)_{p_{0i}, p_{1i}, N_i} \leq \text{var}(Y_i)_{p_{0i}, p_{1i}, N_i}.
\]

and the variance under the convolution model is underdispersed relative to the binomial. We have \( \tau_i^2 = 0 \) when \( N_{0i} = 0 \) or \( N_{1i} = 0 \), or \( p_{0i} = p_{1i} \). Hence, if in a particular table the proportion black is close to zero or one, or the black and white registration probabilities are similar, the variance under the convolution model will be close to the binomial model. We obtain \( \tau_i^2 = 1 \) when \( p_{0i} = 0 \) or 1, and \( p_{1i} = 0 \) or 1. The variance is smaller in the convolution case because we are not sampling \( N_i \) independent Bernoulli random variables each with probability \( q_i \), but, independently, \( N_{0i} \) Bernoulli random variables with probability \( p_{0i} \), and \( N_{1i} \) with probability \( p_{1i} \), a sampling scheme which is more restrictive. If we had sampling with replacement from the table then the binomial model would be appropriate. With fixed numbers in each row the convolution is the correct form (given the other assumptions under which it was developed). There is further discussion of these two models in Section 4.

In the following sections we discuss a number of approaches to ecological inference, and apply these to the Louisiana party registration/race data. Since the internal cells are available for these data, we can determine the bias in the estimated population-averaged rates of Republican registration (a typical quantity of interest in political science applications) which are estimated as:

\[
\hat{\beta}_j = \frac{\sum_{i=1}^{64} N_{ji} \hat{p}_{ji}}{\sum_{i=1}^{64} N_{ji}}.
\]
Table 2. Bias in estimated population-averaged fractions of Republican party registration for the Louisiana data for race $j$, $\hat{p}_j$ (with $p_j$ given by (9)) for a variety of methods; the % bias is given by $100 \times (\hat{p}_j - p_j) / p_j$, $j = 0, 1$.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\hat{p}_0$ (% bias)</th>
<th>$\hat{p}_1$ (% bias)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All individual</td>
<td>0.035 (0)</td>
<td>0.25 (0)</td>
</tr>
<tr>
<td>Ecological regression</td>
<td>0.078 (124)</td>
<td>0.24 (-6.5)</td>
</tr>
<tr>
<td>Nonlinear neighbourhood model</td>
<td>0.14 (303)</td>
<td>0.15 (-41)</td>
</tr>
<tr>
<td>Linear neighbourhood model</td>
<td>0.18 (426)</td>
<td>0.20 (-225)</td>
</tr>
<tr>
<td>Logistic neighbourhood model</td>
<td>0.18 (425)</td>
<td>0.20 (-226)</td>
</tr>
<tr>
<td>Quadratic: $b_0 = 0$, $b_1 = 0.040$</td>
<td>0.060 (73)</td>
<td>0.24 (-3.8)</td>
</tr>
<tr>
<td>Quadratic: $b_0 = -0.040$, $b_1 = 0$</td>
<td>0.086 (149)</td>
<td>0.23 (-7.8)</td>
</tr>
<tr>
<td>Midpoint of bounds</td>
<td>0.23 (565)</td>
<td>0.11 (-596)</td>
</tr>
</tbody>
</table>

where $\hat{p}_{ji}$ are the estimated registration probabilities in area $i$ for $j = 0$ (blacks) and $j = 1$ (whites). The observed values of (9) are given by setting $\hat{p}_{ji} = p_{ji}$, and are given in the first line of Table 2.

Approaches to ecological inference have to some extent been driven by the specific aims in different contexts and so, before reviewing a number of approaches, we briefly compare and contrast the objectives of inference in the social sciences on the one hand, and in epidemiology on the other.

### 3.2. Contrasting Ecological Studies in the Social Sciences and Epidemiology

In this paper the emphasis is on social science applications in which there are large numbers of individuals in each area. Both the response (e.g. Republican registration, committing of a crime) and the predictor (e.g. black/white, unemployed/employed) are typically non-rare (in a statistical sense) and discrete. In epidemiology the response is disease status and is usually rare, and exposures and risk factors may be continuous (for example, the level of exposure to a pollutant or of a dietary variable), or discrete (for example, gender, or a discretised version of a continuous variable such as age).

In political science, although questions of political theory are asked of ecological data, a more typical analysis will examine the differences between racial voting patterns in a specific region. This query can be answered by imputing the number of votes by race for a particular party or candidate, by area. Hence, viewed in this way, prediction rather than causality is the aim. By contrast, in epidemiological applications the usual aim is to estimate the causal effect of an exposure in a certain population and time period (though in public health resource allocation we may want to predict the number of new cases in a future time period, based on the characteristics/demographics of the area). Since the data are usually observational, to estimate the causal effect of the exposure an attempt must be made to control for confounding variables (for example, one or more of age, gender and smoking history) that are responsible for differences in risk, beyond those due to exposure, of the study populations. Hence in ecological studies in epidemiology there is never a single predictor and the data are not simply in the form of a series of $2 \times 2$ tables, since within each area there will (typically) be multiple confounders to control for. Control for confounders in ecological studies is more difficult than in individual studies, since the multivariate distribution of exposures/confounders within and between areas is needed (Richardson et al., 1987, Greenland and Robins, 1994, Prentice and Sheppard, 1995, Plummer and Clayton, 1996, Lasserre et al., 1999).

Although the alleviation of ecological bias using additional variables has been discussed in the social sciences (see, for example the discussion in King, 1997, Section 9.2.1, and Achen and Shively, 1995, Chapter 4), it has been argued (e.g. Flanigan and Zingale, 1985) that using additional predictors changes the inferential question, and so should not be carried out. However, with a model-based approach, imputation of quantities of interest (e.g. voting patterns by race) can still be carried out, by averaging across the additional predictors. This issue of control for additional predictors has some connections to the discussion of population versus individual specific inference in randomized clinical trials with a binary outcomes (e.g. Hauck et al. 1998).
Spatial epidemiological studies are carried out at a variety of geographical scales, depending on the exposure of interest and on data availability, resulting in a variety of population sizes. Inference is required at the level of the individual, but exposures are not generally homogeneous within areas in geographical correlation studies (community-intervention studies provide one counter-example to this). If health outcomes close to a point source of pollution are to be investigated then a small area study is required for there to be any possibility of detection of localised effects with, for example, each area containing a few hundred people (e.g. Wakefield and Elliott 1999). Smaller areas are preferable since the possibility of ecological bias is reduced. In studies of diet and cancer, between-country contrasts are examined since within-country variability in diet is relatively small and within-person variability high (e.g. Prentice and Sheppard, 1990).

As briefly discussed in Section 2, the presence of contextual effects, if not acknowledged, can lead to serious bias. There has been a greater discussion of this aspect in the social sciences than in epidemiology (a notable exception is Greenland, 2001); there are close links between contextual effects and between-area confounding, however.

Perhaps because of the concentration on causal effects rather than imputation, and the need to control for confounders, epidemiology has concentrated more on probabilistic modeling. The rarity of the disease outcome also allows the use of a loglinear model, with its accompanying simplifications; for a non-rare outcome the obvious choice is the logistic model. The latter has received little attention in the social sciences, perhaps because a sampling model for the data is rarely described, rather the mean function is often stated in terms of the discrete predictor variable (as we will describe in Sections 3.4 and 3.5). Similar approaches have been taken in epidemiology with the log of the area-level relative risk being regressed upon area-level exposures, with control for confounders being carried out through the calculation of the expected numbers of cases (e.g. Cook and Pocock, 1983).

Due to the rarity of disease, it is usual to assume that there is no modification of the exposure effect by area in epidemiological studies, in contrast to the social sciences, where hierarchical models have been used for this purpose. If imputation of counts in areas is of interest then allowing, for example, different racial voting patterns by area is extremely important. In ecological regression studies in epidemiology hierarchical models are used to allow the baseline risk to vary by area, possibly with spatially dependent, as well as independent, random effects (such models are discussed in greater detail in Section 5.5).

In epidemiology, due to the difficulties in making causal statements from observational data, there is a long history of carrying out sensitivity studies in order to assess how robust an observed association is to unmeasured confounding, e.g. Cornfield et al. (1959), Rosenbaum and Rubin (1983), and Wakefield (2003) in the context of geographical correlation studies. These approaches are carried out with probabilistic models relating the confounder to the exposure of interest, and the confounder to disease status. Approaches in the same spirit have been considered in the social sciences, but in a less formal manner, see for example Flanigan and Zingale (1985) and Achen and Shively (1995, Chapter 8).

More discussion of the comparison between approaches to ecological inference in epidemiology and the social sciences may be found in Gelman et al. (2001) and Salway and Wakefield (2004).

3.3. Relationship with Simpson’s Paradox

Simpson’s paradox (Simpson, 1951) provides an example of the hazards of making inference from marginal (aggregate) data. Greenland et al. (1999) provide a comprehensive discussion of the paradox, and collapsibility in general, see also Pearl (2000).

Without loss of generality we consider two areas only, and illustrate how an apparent positive association between Republican registration and being white at the ecological level may be the reversal of the association at the level of the individual. Hence, in each of the two areas we assume that at the level of the individual the probability of Republican registration for blacks is greater than that for whites, i.e. $p_{01} > p_{11}$ and $p_{02} > p_{12}$, and $x_1 < x_2$ so that area 1 has a greater proportion white than area 2. The ecological fallacy occurs when the overall proportion with Republican registration is greater in area 1, even though this area
has a greater white population, since then it appears that being white and being registered as Republican are positively associated. Recalling that $Y = 1$ corresponds to Republican registration, the fallacy may be expressed mathematically as

$$q_1 = \Pr(Y = 1|\text{area 1}) = \sum_{j=0}^{1} \Pr(Y = 1|\text{area 1}, \text{race } j) \Pr(\text{race } j|\text{area 1}) = p_{01}x_1 + p_{11}(1-x_1) > p_{02}x_2 + p_{12}(1-x_2) \quad (10)$$

$$q_2 = \sum_{j=0}^{1} \Pr(Y = 1|\text{area 2}, \text{race } j) \Pr(\text{race } j|\text{area 2}) = \Pr(Y = 1|\text{area 2}) = q_2$$

so that the aggregate conclusion is that the probability of Republican registration for blacks is greater in area 1 than in area 2, leading to the ecological fallacy.

For the inequality in (10) to occur, area must be acting as a confounder, that is Republican registration must be associated with area, and the proportion of blacks must differ across areas. The latter is usual, and the former may occur for a variety of reasons including a contextual effect of race, or confounders associated with the area such as income. Here we are assuming that area does not lie on the causal pathway between race and Republican registration (as may happen when re-districting is influenced by racial breakdown).

As an example we take two areas from the Louisiana data for which the proportions white were $x_1 = 0.16$ and $x_2 = 0.51$, and the proportions registered 0.31 and 0.08, respectively. Then, for example, if $p_{01} = 0.35$, $p_{11} = 0.30$, $p_{02} = 0.12$, $p_{12} = 0.08$, so that the blacks are more likely to have Republican registration in each area, we obtain

$$q_1 = 0.35 \times 0.16 + 0.30 \times 0.84 = 0.31$$

$$q_2 = 0.09 \times 0.51 + 0.07 \times 0.49 = 0.08.$$ 

The fallacy/paradox occurs here because in area 1 the probabilities of Republican registration are much higher than in area 2 for both whites and blacks, and there are a greater number of whites in this area.

A typical formulation of Simpson’s paradox is the following. We observe the entries of a $2 \times 2$ table in which treatment A has a higher recovery rate than treatment B. However, if we had observed the data as a series of $2 \times 2$ tables representing treatment $\times$ recovery for each of $m$ age groups, then treatment B would provide a higher recovery rate than treatment A for each age group. In the ecological setting we observe the race $\times$ registration margins in each of $m$ areas. Hence the two scenarios are not identical; in the conventional setting we have a data reduction from $2m$ pieces of information to just 2, while in the ecological setting the reduction is to $m$ quantities, with the margin of interest being collapsed over.

In summary, there are two aspects of marginalisation that are relevant in ecological inference: firstly, in the presence of a confounder (area) the effect of interest may reverse sign when viewed from individual and aggregate perspectives; and secondly, effect measures are not consistent across levels of aggregation even without confounding, so even in the absence of contextual effects “bias” may be present in the estimates of race-specific fractions of Republican registration; here we place bias in quotes because bias is relative to the quantity that is being estimated and in some situations we may be interested in the marginal association.

### 3.4. Ecological Regression

Goodman (1953, 1953) provided one of the earliest and clearest discussions of ecological inference, and discussed a method, a crucial assumption of which is that

$$E[p_{0|i}|x_i] = \tilde{p}_0, \quad E[p_{1|i}|x_i] = \tilde{p}_1,$$

for all $i$ so that the expectations of the black and white registration fractions in an area do not depend on the race proportions in that area (where the expectations of the race-specific fractions are taken over
hypothetical areas with the same black/white proportions). Under these assumptions we obtain

$$E[\hat{q}_i | x_i] = \alpha + \beta x_i,$$

where $\alpha = \hat{p}_1$, $\beta = \hat{p}_0 - \hat{p}_1$. Hence when $\hat{q}_i = y_i/N_i$ is regressed upon $x_i$ we obtain estimates $\hat{\alpha}$ and $\hat{\beta}$, and without further assumptions (such as the distribution of the deviations of $\hat{p}_{ji}$ from their expectations) we obtain estimates $\hat{p}_0 = \hat{\alpha} + \hat{\beta}, \hat{p}_1 = \hat{\alpha}$, which do not depend on $i$.

There are a number of difficulties with this method. The most serious drawbacks are that there is no allowance for contextual effects, and it is unrealistic to assume no heterogeneity in registration fractions across counties, we would expect demographic and county characteristics to modify these fractions. A less serious problem is that in its original form least squares, which does not acknowledge the non-constant variance, was used for estimation; estimates of the fractions are also not constrained to lie within their admissible ranges of $(0,1)$. The latter has been taken to be evidence of the inappropriateness of the model, see for example Gelman et al. (2001).

Goodman (1959) was aware of the above drawbacks and in particular questioned the appropriateness of the constancy assumption in many situations, and also discussed: the non-constancy of variance in the observed marginal outcome probability; the introduction of area-level covariates; interactions with area-level characteristics; and the search for clusters of areas with probabilities displaying similar behaviour. In spite of these problems the method has become popular, and is often referred to as Goodman regression; because Goodman did not universally endorse the approach, we prefer the name ecological regression.

Figure 3 shows the weighted least squares line for the Louisiana registration/race data. The black and white fractions are estimated by the points at which the regression line intersects the $x = 1$ and $x = 0$ lines, respectively, emphasizing that the estimates and $x$ are uncorrelated. Table 2 provides summaries corresponding to the use of model (11), and the area-race specific registration fractions are plotted in Figure 4(b). In Figure 4(a) we provide the true fractions, which correspond to Figures 2(c) and (d) combined. Under ecological regression there is overestimation of $\hat{p}_0$ and underestimation of $\hat{p}_1$, though for the white fractions a reasonably accurate estimate is obtained. From Section 2 we would not expect this method to be accurate since there are both considerable between-area differences in the race-specific registration fractions, particularly for whites, and contextual effects.

3.5. Further Regression Methods

The so-called “nonlinear neighbourhood model” (Freedman et al., 1991), assumes that $p_{0i} = p_{1i} = q_i$, so that race and registration are independent at the individual level and do not vary with the proportion black, but are heterogeneous across areas. Since the probabilities within each table are equal, the table may be collapsed and estimation is straightforward, giving $\hat{p}_{0i} = \hat{p}_{1i} = \hat{q}_i = y_i/N_i$. Freedman (1999) states that in this model, “...behaviour is determined by geography not demography”. Though this model acknowledges that the probabilities of registration vary across areas, it will not be realistic in many situations. An alternative, the linear neighbourhood model, again assumes that the black and white probabilities are equal, but assumes that they both depend on the proportion black so that $E[p_{0i} | x_i] = E[p_{1i} | x_i] = \alpha + \beta x_i$, and so contextual effects determine an area’s fractions, and an individuals race is again irrelevant. This model leads to $E[q_i | x_i] = \alpha + \beta x_i$, which is indistinguishable from (11) (and was the motivation for Freedman et al. (1991) to discuss the model); the estimates of the area-race specific fractions are vastly different, however. Figures 4(c) and (d) give the estimates under the linear and nonlinear neighbourhood models, respectively. The differences between the estimates under the linear neighbourhood and ecological regression models (panels (c) and (b), respectively) is striking.

Both the nonlinear and the linear neighbourhood models give identical estimates of the total number of blacks and whites registered as Republican. Gelman et al. (2001) note that the neighbourhood models are valid for at most one level of aggregation in the sense that if, for example, we combine two areas for which $p_{0i} = p_{1i} = q_i$ holds, then for the combined area this can no longer, in general, be true. This aspect has similarities with the introduction of normal random effects in spatial aggregate studies; since the sum of
Fig. 4. For the Louisiana registration/race data, estimates of fractions registered Republican for blacks (△) and whites (▽) versus proportion black, in a variety of scenarios. (a) True fractions. (b) Ecological regression. (c) Linear neighbourhood model. (d) Nonlinear neighbourhood model. (e) Quadratic regression model, contextual effect for whites only. (f) Quadratic regression model, contextual effect for blacks only.
lognormals is not lognormal, this model can only be true at one level of aggregation (Wolpert and Ickstadt, 1998); we do not view this as a fatal flaw, however, since this aspect alone does not invalidate the model’s usefulness. Table 2 confirms that, as expected in this context, the neighbourhood models provide very poor estimates for the Louisiana data (we saw in Section 2 that the race of an individual is strongly associated with registration).

A number of authors (e.g. Goodman, 1959, Achen and Shively, 1995) have assumed that

$$p_{ji} = a_j + b_j x_i,$$  \hspace{1cm} (12)

so that area-specific fractions depend on both the race of the individual, and upon contextual race effects. In this case we obtain

$$q_i = \alpha + \beta x_i + \gamma x_i^2,$$

where $$\alpha = a_1$$, $$\beta = a_0 + b_1 - a_1$$ and $$\gamma = b_0 - b_1$$ so that we have four unknowns to estimate from three parameters. If we assume that only one of the race-specific probabilities depends on a contextual effect, that is $$b_0 = 0$$ or $$b_1 = 0$$, then all parameters are identifiable. If $$b_0 = 0$$ we have contextual effects for whites only with estimates

$$p_{0i} = \hat{a}_0 + \hat{a}_1 x_i = \hat{a}_0 + \hat{\beta} x_i,$$

and if $$b_1 = 0$$

$$p_{0i} = \hat{a}_0 + \hat{b}_0 x_i = \hat{a}_0 + \hat{\beta} + \hat{\gamma} x_i$$

and

$$p_{1i} = \hat{a}_1 = \hat{\alpha}.$$

In both cases $$\hat{\gamma}$$ gives an estimate of the linear contextual effect.

When the quadratic model is fitted to the Louisiana data using ordinary least squares we obtain an estimate (standard errors) of the quadratic term of $$\hat{\gamma} = -0.09 (0.42)$$; with weighted least squares we obtain $$\hat{\gamma} = -0.04 (0.44)$$. Residuals were inspected from the linear model and showed little evidence of curvature (as confirmed by Figure 3), which is consistent with the estimate of the quadratic term and the associated standard errors. The quadratic term produces substantial changes in imputed cells over the linear model, however, the estimates from the latter quadratic fits are shown in panels (e) and (f) of Figure 4. Table 2 show that the quadratic models provide improved estimates over ecological regression with a contextual effect for the white fractions but poorer estimation with the contextual effect for the black fractions (we saw in Figure 2(c) and (d) that the more dominant contextual effect was for whites). There is no way of knowing from the data alone which of the two contextual forms is the appropriate one. This exercise has illustrated the sensitivity of inference to modeling assumptions that are uncheckable from the data alone. We finally note that the sensitivity of inference to the specific quadratic model can be examined further by fixing one of $$b_0$$, $$b_1$$ to plausible values other than zero.

Chambers and Steel (2001) consider methods for avoiding the linearity assumption and for the modelling of contextual effects. One of their suggested approaches replaces the linear model by a local linear smoother, with estimation following from an estimating function; additional covariates may be incorporated and bounds are enforced to produce admissible estimates. In a second method a simple semi-parametric approach is taken in which it is assumed that $$E[p_{1i}|x_i] = \theta E[q_i|x_i]$$, in area i. The unknown $$\theta$$ is constrained in order to keep $$p_{0i}$$ and $$p_{1i}$$ in their correct ranges. This model is asymmetric in the sense that it implies that

$$E[p_{0i}|x_i] = \frac{E[q_i|x_i](1 - \theta(1 - x_i))}{x_i},$$

As with the neighbourhood model, this model is clearly unrealistic in some situations but provides accurate estimates in the examples of Chambers and Steel (2001) for which $$p_{0i}$$ and $$p_{1i}$$ are similar. Thomsen (1987) provides an excellent and in-depth discussion of electoral transitions, and proposes a model based on a latent variable approach; for discussion of this model see Cleave et al. (1995) and King (1997).

### 3.6. General Formulation

The exact correspondence between ecological regression and the linear neighbourhood model and the lack of identifiability of the parameters in the quadratic model, arise because the contextual effects are assumed
linear in \( x \). This model is not natural from a statistical perspective, however, since it can provide fitted probabilities outside of \([0,1]\). A general model that allows registration probabilities to depend on both the race of the individual and the contextual effects of race (with this effect assumed constant across areas) is given by

\[
p_{ji} = \Pr(Y = 1 \mid \text{race } j, \text{ area } i) = g^{-1}(\alpha_{0i} + \beta_0 x_i),
\]

for a link function \( g(\cdot) \), leading to the marginal mean

\[
q_i = \Pr(Y = 1 \mid \text{area } i) = g^{-1}(\alpha_{0i} + \beta_0 x_i) + g^{-1}(\alpha_{1i} + \beta_1 x_i)(1 - x_i).
\]

The ecological regression method corresponds to this model with a linear link and \( \alpha_{0i} = \alpha_0, \alpha_{1i} = \alpha_1 \) and \( \beta_0 = \beta_1 = 0 \), the nonlinear neighbourhood model has \( \alpha_{0i} = \alpha_{1i} = \alpha_1 \), and \( \beta_0 = \beta_1 = 0 \), the linear neighbourhood model has \( \alpha_{0i} = \alpha_{1i} = \alpha \) and \( \beta_0 = \beta_1 = \beta \) with a linear link, and the quadratic model has a linear link with \( \alpha_{0i} = \alpha_0 \) and \( \alpha_{1i} = \alpha_1 \). The local linear method of Chambers and Steel (2001) may be viewed within this framework with a link function that varies with \( x_i \).

Hence it would appear profitable to consider the general form with a non-linear link, logistic and probit forms being the obvious choices. In these cases the model with individual and contextual effects has all parameters identifiable. Unfortunately assuming a nonlinear link theoretically removes the non-identifiability but in practice contextual effects are often not highly unstable. This issue has close connections with criticisms made of a class of methods for avoiding selection bias, for example that proposed by Heckman (1979); see for example Little (1985) and Copas and Li (1997). As Achen and Shively (1995, p. 117) point out, since in practice contextual effects are often not strong, it is virtually impossible to distinguish between linear and nonlinear forms (such as logistic or probit models).

To illustrate this indistinguishability we fit a linear logistic neighbourhood model in which \( p_{0i} = p_{1i} = q_i \) and \( \log \left( \frac{q_i}{1 - q_i} \right) = \alpha + \beta x_i \), and with \( Y_i | q_i \), assumed to be binomially distributed. The population-averaged estimates are reported in Table 2 and are virtually identical to those from the linear neighbourhood model. This was confirmed by plotting the estimated fractions by area from the two models. This illustrates that it is not possible to distinguish between linear and logistic forms from the data alone.

### 3.7. Method of Bounds

Duncan and Davis (1953) noted that bounds could be placed on quantities of interest by examining the admissible values of \( Y_{0i}, Y_{1i} \), conditional on the margins. From (5), since \( l_i \leq Y_{0i} \leq u_i \), we have the following bounds on the fraction \( Y_{0i}/N_{0i} \):

\[
\max \left\{ 0, \frac{\tilde{q}_i - (1 - x_i)}{x_i} \right\} = \max \left\{ 0, \frac{Y_i + N_{0i} - N_i}{N_{0i}} \right\} \leq \tilde{p}_{0i} \leq \min \left\{ 1, \frac{Y_i}{N_{0i}} \right\} = \min \left\{ 1, \frac{\tilde{q}_i}{x_i} \right\}
\]

and, similarly for \( Y_{1i}/N_{1i} \):

\[
\max \left\{ 0, \frac{\tilde{q}_i - x_i}{1 - x_i} \right\} = \max \left\{ 0, \frac{Y_i + N_{1i} - N_i}{N_{1i}} \right\} \leq \tilde{p}_{1i} \leq \min \left\{ 1, \frac{Y_i}{N_{1i}} \right\} = \min \left\{ 1, \frac{\tilde{q}_i}{1 - x_i} \right\}
\]

\( i = 1, \ldots, m \). This development makes it clear that the bounds are implicit in the convolution likelihood (4) through the range on permissible values of the internal cells. The bounds are on the true fractions of individuals \( Y_{ji}/N_{ji} \), and not on \( p_{ji} \); for any set of table margins, any values \( 0 < p_{ji} < 1 \) are theoretically possible. King (1997, p. 77) amongst others notes that the bounds are often too large to be useful. Although examination of the bounds is a useful exploratory measure, a statistical model is preferable for inference since sources of variability can be formally acknowledged.

Table 3 lists the different types of bounds that are possible, depending on the values of \( x_i = N_{0i}/N_i \) and \( \tilde{q}_i = y_i/N_i \) observed in a particular table. These cases are illustrated in Figure 4(a). For any table we obtain at most two informative restrictions on \( \tilde{p}_{0i}, \tilde{p}_{1i} \), and it is not possible for the lower bound of one estimate
to be above zero, and the upper bound for another estimate to be below one, simultaneously so that the bound on $\hat{\pi}_{1i} - \hat{\pi}_{0i}$ never excludes zero (illustrating the universal potential for the ecological fallacy). This is another indicator of the lack of information in the margins concerning the difference. Table 3 also gives the bounds on the difference of the fractions $\hat{\pi}_{1i} - \hat{\pi}_{0i}$; the “best” potential scenario, in terms of the narrowness of the bounds corresponds to cases 1 and 4 with $\hat{q}_i$ close to zero or close to one, respectively. The limiting situations of $\hat{q}_i = 0$ (case 1) or $\hat{q}_i = 1$ (case 4) produce zero-width bounds for both fractions, but there is obviously little information to discriminate between $p_{0i}$ and $p_{1i}$ in this situation. If $x_i = 0$ then the bounds on $\hat{p}_{0i}$ are $(0,1)$ and the lower and upper bounds equal $\hat{q}_i$ for $\hat{p}_{1i}$; the case $x_i = 1$ follows in a symmetric fashion. In these cases the width of the bound on $\hat{p}_{1i} - \hat{p}_{0i}$ is 1. Over all tables, the information that is available depends crucially on the range of observed $x_i$’s, that is, the ‘design’. As we discuss later, we would ideally like to have areas with $x$ close to both 0 and 1, so that we can have informative bounds for both $\hat{p}_{0i}$ and $\hat{p}_{1i}$. An extremely important observation is that for a given $x_i$, the bounds will not, in expectation, be centred on the true values of the fractions.

In epidemiological applications we are in case 1, with $\hat{q}_i$ typically much smaller than $x_i$ and $1 - x_i$. Hence we obtain upper bounds on disease risk in area $i$, for exposed and unexposed individuals, but these bounds will not be useful for a rare disease. This is the same situation as highlighted by Selvin (1958) for the often-quoted example of Durkheim’s suicide (Durkheim, 1897), in which suicide rates were examined as a function of the proportion of Protestant residents.

Figure 5(b)–(d) show the bounds on the fractions with Republican registration for the Louisiana data. The plots in panels (b) and (c) are routinely used in the examples of King (1997). For these data, as expected, due to the distribution of $x_i$ across areas, the bounds for $\hat{p}_{1i}$ are more informative than the bounds for $\hat{p}_{0i}$, the average width of the bounds is 0.57 for $\hat{p}_{0i}$ and 0.19 for $\hat{p}_{1i}$ showing that there is greater information for the latter since $\hat{q}_i < x_i$ in the majority of counties and $\hat{q}_i < 1 - x_i$ in all counties (case 1 in Table 3, Figure 5(a)) and so we have the upper bound $\hat{p}_{1i} < 1$ in all counties. The bounds on $\hat{p}_{0i}$ become narrower with increasing $x_i$, again as expected. In Figure 5(d) we see that for values of $x$ close to zero the bounds on the difference in fractions suggests that $\hat{p}_{1i} > \hat{p}_{0i}$. From the individual-level data we know that the latter conclusion is incorrect, and stems from the lack of information on $\hat{p}_{0i}$. The dangers of making inference from the bounds alone are also revealed as $x$ increases; closer to 0.5 the bounds on the difference become more symmetric about 0.

### Table 3. Possible bound types, $l_{0i} = (\hat{q}_i - 1 + x_i)/x_i > 0$, $l_{1i} = (\hat{q}_i - x_i)/(1 - x_i) > 0$, $u_{0i} = \hat{q}_i/x_i < 1$, $u_{1i} = \hat{q}_i/(1 - x_i) < 1$.

<table>
<thead>
<tr>
<th>Case</th>
<th>Relationships</th>
<th>$p_{0i}$ bound</th>
<th>$p_{1i}$ bound</th>
<th>$\hat{p}<em>{1i} - \hat{p}</em>{0i}$ bound</th>
<th>Width of bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$0 &lt; \hat{q}_i \leq x_i &lt; 1$</td>
<td>$(0, u_{0i})$</td>
<td>$(0, u_{1i})$</td>
<td>$( -u_{0i}, u_{1i})$</td>
<td>$\hat{q}_i/(x_i(1 - x_i))$</td>
</tr>
<tr>
<td>2</td>
<td>$0 &lt; 1 - x_i \leq \hat{q}_i &lt; x_i &lt; 1$</td>
<td>$(u_{0i}, 1)$</td>
<td>$(0, 1)$</td>
<td>$(-u_{0i}, 1 - x_i)$</td>
<td>$(1 - x_i)/(x_i)$</td>
</tr>
<tr>
<td>3</td>
<td>$0 &lt; x_i \leq \hat{q}_i \leq 1 - x_i &lt; 1$</td>
<td>$(0, 1)$</td>
<td>$(l_{1i}, u_{1i})$</td>
<td>$(l_{1i} - 1, u_{1i})$</td>
<td>$1 + x_i/(1 - x_i)$</td>
</tr>
<tr>
<td>4</td>
<td>$0 &lt; x_i \leq \hat{q}_i &lt; 1$, $0 &lt; 1 - x_i \leq \hat{q}_i &lt; 1$</td>
<td>$(l_{0i}, 1)$</td>
<td>$(l_{1i}, 1)$</td>
<td>$(l_{1i} - 1, 1 - l_{0i})$</td>
<td>$(1 - \hat{q}_i)/(x_i(1 - x_i))$</td>
</tr>
</tbody>
</table>

4. **A Likelihood Approximation**

In this section we discuss a plausible likelihood in the absence of contextual effects. With the assumptions of race-specific probabilities in each area, the likelihood takes the form of a convolution, as given in equation (4).

This convolution is computationally intensive to calculate for the area sizes that are typical in social science applications (see the Appendix for more discussion), but a convenient normal approximation is available. We do not assume constant probabilities across areas and approximate each of the two binomial likelihoods (corresponding to each row of the 2 × 2 table) by normal distributions, to give the likelihood in area $i$ as

$$Y_i|p_{0i}, p_{1i} \sim \int \phi \left( \frac{y_{0i} - N_{0i}p_{0i}}{\sqrt{N_{0i}p_{0i}(1 - p_{0i})}} \right) \phi \left( \frac{Y_i - y_{0i} - N_{1i}p_{1i}}{\sqrt{N_{1i}p_{1i}(1 - p_{1i})}} \right) dy_{0i},$$
Fig. 5. (a) Information in bounds as a function of $x$ and $q$ (see Table 3 for description of the different cases). For the Louisiana registration/race data, deterministic bounds on: (b) $\tilde{p}_{0i}$, (c) $\tilde{p}_{1i}$, (d) $\tilde{p}_{1i} - \tilde{p}_{0i}$, $i = 1, ..., 64$. 
where \( \phi(x) \) denotes the density of a standard normal random variable. This approximation may be justified using a local limit theorem (e.g., Grimmett and Stirzaker, 1992, p.177), and leads to

\[
Y \mid p_{0i}, p_{1i} \sim N\{\mu_1(p_{0i}, p_{1i}), V_1(p_{0i}, p_{1i})\}.
\]

where

\[
\mu_1(p_{0i}, p_{1i}) = N_0q_1 = N_0p_{0i} + N_1p_{1i}
\]

and

\[
V_1(p_{0i}, p_{1i}) = N_0p_{0i}(1 - p_{0i}) + N_1p_{1i}(1 - p_{1i}).
\]

At first sight this approximation would seem to simply approximate the sampling distribution of \( Y \mid p_{0i}, p_{1i} \) by a normal with mean and variance given by the marginal moments, but we emphasize that the count in each row is being approximated by a normal, and the accuracy of these approximations is determined by the size of the row totals, and by the closeness of \( p_{0i} \) and \( p_{1i} \) to zero or one. Unfortunately in an ecological setting, due to the non-identifiability, the data alone do not allow these extreme cases to be ruled out. However, the approach we advocate is the use of this likelihood within a hierarchical setting and so we only require an accurate likelihood approximation in those regions supported by the prior. Further investigation of the above approximation, and other possibilities, perhaps including continuity corrections, would be desirable, particularly to determine the accuracy as a function of \( N_0i, N_1i \) and \( p_{0i}, p_{1i} \).

We now briefly examine the shape of the likelihood for a single table, through the approximation given by (15), which yields the log-likelihood

\[
L(p_{0i}, p_{1i}) = -\frac{1}{2} \log V_1(p_{0i}, p_{1i}) - \frac{(Y_i - \mu_1(p_{0i}, p_{1i}))^2}{2V_1(p_{0i}, p_{1i})}.
\]

To determine the maximum likelihood estimator (MLE) we reparameterise in terms of \( \mu_i = \mu_1(p_{0i}, p_{1i}) \) and \( V_i = V_1(p_{0i}, p_{1i}) \) and consider the profile likelihood. Both \( \mu_i \) and \( V_i \) are defined between 0 and \( N_i \). The likelihood is maximized by \( \hat{\mu}_i = Y_i \), independent of \( V_i \), and the profile likelihood for \( V_i \) is given by the first term in (17) and hence we need to minimize (16). This variance is a quadratic surface in \( p_{0i}, p_{1i} \) with maximum at \((\hat{q}_i, \hat{q}_i)\) and is minimized at one of the endpoints of the “tomography line” (King, 1997, Gelman et al. 2001), the latter being given by

\[
\hat{p}_{1i} = \frac{\hat{q}_i}{1 - x_i} = \frac{x_i}{1 - x_i}.
\]

In fact the minimizing point is closest to one of the corners so that the MLE lies at one endpoint. Consideration of the constraints on \( p_{0i}, \hat{p}_{1i} \) reveals which endpoint; the MLEs are given in Table 4. For each of the four corners there are two possibilities depending on which of the lines joining at that corner is crossed by the tomography line. For example, if the tomography line is closest to the \((0,0)\) corner at the point at which it crosses the \( p_{0i} = 0 \) line then the MLE is \((0, u_{1i})\) whereas if the line crosses the \( p_{1i} = 0 \) line then the MLE is \((u_{0i}, 0)\). These points are useful for determining the information in the likelihood alone, we discuss this further in Section 6. Over all tables, the asymmetry in the heights of the likelihoods at the two endpoints is a factor in the bias that can be expected when hierarchical models are applied.

Plackett (1977) showed that the maximum likelihood estimate of the log odds ratio in a single table in which the margins only are observed is \( \pm \infty \), which corresponds to \( p_{0i} = 0 \) or 1 and/or \( p_{1i} = 0 \) or 1. Steel, Beh and Chambers (2004) work with the convolution directly and report that the maximum likelihood estimator lies at the endpoint of the tomography line.

The binomial likelihood of King et al. (1999) is constant along the tomography line. Figure 6 shows (a) the convolution, (b) the normal approximation, and (c) the binomial likelihood, in each case the likelihoods have been scaled to have maximum 1. We see that the normal approximation retains the features of the convolution, while the binomial is constant along the tomography line. For these data we have \( N_i = 70, x_i = 2/7 \) and \( \hat{q}_i = 3/7 \) so that we are in case 3 of Table 3 with bounds \((0.1)\) for \( p_{0i} \) and \((0.2, 0.6)\) for \( p_{1i} \). We see that there is appreciable mass lying away from the tomography line. The tomography line goes closest
to the corner (1,0), crossing the $p_0 = 1$ line and hence, from Table 4, the MLEs of $(\hat{p}_{0i}, \hat{p}_{1i})$ are at (1,0.2), as confirmed by Figures 6(a) and (b).

The normal approximation (17) does not involve any bounds on $p_{0i}$, $p_{1i}$ (beyond the usual restrictions to (0,1)), but this is not a problem since it is the fractions $\hat{p}_{0i}, \hat{p}_{1i}$ that are constrained by the overall fraction $\hat{q}_i$. As $N_i \to \infty$ with $x_i$ and $\hat{q}_i$ constant, the convolution likelihood function (and the normal approximation) becomes concentrated along the tomography line (3) with an asymmetric U-shape due to the quadratic form of $V_i(p_{0i}, p_{1i})$ in (16), with the maximum at one endpoint. The truncated normal model of King (1997), which we discuss in Section 5.2 (equation (19)), implicitly assumes that the first stage is constant along the tomography line, and zero elsewhere.

Hawkes (1969) and Brown and Payne (1986) use a normal approximation for $(Y_1, \ldots, Y_m)$, without specifically considering the constituent normal approximations for each of the rows, see also Achen and Shively (1995, p. 47) who gave the normal form for the ecological regression model.

It is clear that the data in one table alone give very little information concerning $p_{0i}, p_{1i}$. The method of bounds is often referred to as “deterministic” (King, 1997; Cho, 1998) and if we were to examine the likelihood for each table alone with a uniform prior on each of $p_{1i}$ then essentially this would be a stochastic version of this method (and the midpoint of the bounds will approximate the posterior median under a uniform prior).

Since the bounds are highly dependent on the particular $x$ value of the table under consideration, it is clear that the midpoint of the bounds will in general be highly biased, even in the absence of contextual effects. Table 2 shows that using the midpoint of the bounds provides very poor inference. For $p_{0i}$ the bounds are (0,1) for many areas (with $x_i < \hat{q}_i < 1 - x$, case 3 of Table 3) and the midpoint is 0.5 in this case (Figure 6(b)). This overestimation also causes underestimation of $p_{1i}$ because of the dependence between $p_{0i}$ and $p_{1i}$, summarized by (3).

## 5. Hierarchical Models

### 5.1. Introduction

Hawkes (1969) considered the $R \times C$ table case, derived the mean and covariance of the marginal data and proposed a number of models and estimation methods, including a random effects model that was not implemented due to the difficulty at that time of the required computations. Brown and Payne (1986) also examined the general $R \times C$ case and modelled the aggregate data using a convolution of dirichlet-multinomial distributions, which they referred to as an aggregated compound multinomial (ACM) distribution. Inference was carried out via maximum likelihood estimation by obtaining the marginal moments for all $m$ tables, and then assuming a multivariate normal distribution for the observed marginal response, thus avoiding the calculation of the convolution likelihood; this numerical maximization is difficult, however. Transforms of the raw counts via the Box-Cox family were used to improve the accuracy of the multivariate normality.
Fig. 6. Likelihood functions for a table with row margins $N_0 = 20$ and $N_1 = 20$, $N_i = 70$, and $Y_i = 30$ (so that $x_i = 2 = 7$ and $q_i = 3 = 7$): (a) convolution, (b) normal approximation, (c) binomial.
assumption, and area-level covariates were naturally incorporated via a loglinear model.

A hierarchical approach in the spirit of Section 3.5 would be the model

\[ p_{ji} = g^{-1}(\alpha_{ji} + \beta_{j}x_{i}). \]

The choice of a probit link \( g(\cdot) \), combined with \( \alpha_{ji} \sim N(\alpha_{j}, \sigma^{2}_{j}) \), yields a closed form mean function

\[ q_{i} = \Phi \left( \frac{\alpha_{0} + \beta_{0}x_{i}}{1 + \sigma^{2}_{0}} \right) x_{i} + \Phi \left( \frac{\alpha_{1} + \beta_{1}x_{i}}{1 + \sigma^{2}_{1}} \right) (1 - x_{i}), \]

and results in a likelihood that depends on six parameters. Unfortunately this model also suffers from instability problems, without strong priors on the hyperparameters, and is highly dependent on the specific link/linear model chosen, as discussed in Section 3.6.

The appeal of hierarchical models is that they allow the data from all areas to combine to indicate where in a particular table the probabilities lie along the tomography line.

### 5.2. Truncated Normal Prior

In a significant contribution, King (1997) combined aspects of the method of bounds and ecological regression by keeping distinct pairs of fractions and incorporating into the estimation procedure the \( m \) tomography lines given by

\[ \tilde{p}_{1i} = \frac{\tilde{q}_{i} - \tilde{p}_{0i} x_{i}}{1 - x_{i}}, \]

which correspond to a rearrangement of (3) with the fractions \( \tilde{q}_{i}, \tilde{p}_{0i}, \tilde{p}_{1i} \) substituted for the population probabilities \( q_{i}, p_{0i}, p_{1i} \), hence incorporating the bounds.

In the likelihood approaches described so far we have considered the table/race-specific probabilities of Republican registration \( p_{ji} \). King models the fractions \( \tilde{p}_{ji} \) (see also the footnote of Cho, 1998, p. 155), which would be given by \( y_{ji}/N_{ji} \), if the internal cells were observed. King (1997) imposes (18) by implicitly assuming a (first-stage) “likelihood” of

\[ L(\tilde{p}_{0i}, \tilde{p}_{1i}) = \begin{cases} 
1 & \text{if } \tilde{q}_{i} = \tilde{p}_{0i} x_{i} + \tilde{p}_{1i} (1 - x_{i}) \\
0 & \text{otherwise.}
\end{cases} \]

where \( \tilde{q}_{i} = y_{ji}/N_{i} \), see King (1997), equation (7.1). Hence the manner in which the data enter the model is via the accounting equation (18), and the “likelihood” is flat along this line. We use the term likelihood loosely here to refer to the missing data stage of the model. In this formulation, there are no hypothetical probabilities arising from an infinite population, but rather the fact that each unobserved fraction lies along the tomography line is being imposed, and it is these fractions that are viewed as exchangeable and arising from a second stage distribution. As discussed in Section 4, given all of the other difficulties of ecological inference, having a likelihood that is constant along the tomography line will often be of secondary importance in tables with large \( N_{i} \), since examination of the convolution likelihood along the tomography line reveals that the likelihood does not vary greatly since there is so little information in the data to discriminate between \( \tilde{p}_{0i} \) and \( \tilde{p}_{1i} \), beyond the bounds. The rationale for King’s approach is that the bounds can be tightened once the bound for one fraction is known since the fractions are constrained to lie on the tomography line. The approach may also be seen as a limiting case of the binomial likelihood since as \( N_{i} \rightarrow \infty \) the latter tends to a flat line corresponding to the tomography line (Figure 6(c)).

To overcome the intrinsic lack of information, and lack of identifiability of the fractions, King (1997) assumes that the pair \( \tilde{p}_{0i}, \tilde{p}_{1i} \) arise as an independent and identically distributed sample from a truncated bivariate normal distribution. In this way, constant fractions across tables are avoided, but the hierarchical model allows different areas to borrow strength from each other. King (1997) regards the truncated normal constrained to lie along the tomography line as the likelihood for the data \( \tilde{q}_{i} \). This model is controversial (e.g., Freedman et al., 1998, 1999; King, 1999; Cho, 1998) because the model, in its original form, assumes
that $p_{0i}$ and $p_{1i}$ are independent of $x_i$, i.e. that there are no contextual effects. Secondary worries are the assumption of a truncated normal distribution (and the difficulty in determining whether this assumption is reasonable), and the specification of third stage priors, if used (this point is considered in greater detail in Section 6). Area-level characteristics (including $x_i$) can, in principle, be incorporated via a regression model (see King, 1997, p. 180) but the form of the relationship is difficult to determine and the amount of information that can be extracted on this relationship is typically small and prior information is essential (this is demonstrated in Section 7).

King (1997) provides a number of diagnostic procedures for assessing the assumptions of the model including examination of the tomography plot to see if there are multiple points of intersection; and plots of the bounds versus $x_i$ (Figure 5(b), (c)). The merit of these diagnostics has been questioned, however (Freedman et al., 1998, 1999; Cho, 1998). Gelman et al. (2001) consider a number of diagnostics including the examination of binned residuals plotted versus $x_i$, and interpreting the situation in which the estimates from the ecological regression are inadmissible.

In this paper a hierarchical modelling approach is advocated but one cannot expect such an approach to be reliable in all situations, and sensitivity analyses, in particular to the existence and strength of contextual effects, is to be encouraged. A major problem is in determining the appropriateness of a particular hierarchical model, regardless of the number of areas and the size of the table margins.

5.3. Beta Priors

King et al. (1999) use the binomial likelihood (6) and at the second stage assume that $p_{0i}$ and $p_{1i}$ are independent with

$$p_{ji}|a_j, b_j \sim \text{Be}(a_j, b_j),$$  \hspace{1cm} (20)

for $i = 1, ..., m$, and with independent exponential hyperpriors $\text{Exp}(\lambda_j)$ on $a_j, b_j$, $j = 0, 1$, at the third stage, where $\text{Exp}(\lambda)$ denotes an exponential distribution with mean $\lambda^{-1}$. King et al. (1999) pick $\lambda = 0.5$ in their application, labelling this “a fairly non-informative distribution at the final level”, but this will often be a poor choice since it places too little prior weight on larger values of $a_0, a_1, b_0, b_1$. It is important to allow large values of these parameters because such values provide a more informative prior (since $a_j + b_j - 2$ behaves like a sample size), to supplement the first stage information for race $j$.

Another important consideration is the marginal prior distribution for $p$ that results when we integrate over stages 2 and 3 of the model. Figures 7(a) and (b) illustrate, respectively, the marginal prior for $p_{ji}$, after integration over $a_j, b_j$, under the $\text{Exp}(0.5)$ hyperprior, and the choice $\text{Exp}(0.01)$ which will be used for comparison in Sections 6 and 7. With mean $\lambda^{-1} = 2$ hyperpriors the distribution is highly U-shaped which will often be undesirable. In King et al. (1999, Figures 3 and 4) a number of cases of bimodality in the posterior were reported and it seems likely that these were due to this extreme form of prior.

An alternative prior choice places a uniform prior on the mean $\phi = a/(a+b)$ and a distribution on a parameter determining the scale $\gamma = a + b$. Kahn and Raftery (1996) use such a prior with the improper choice $p(\gamma) \propto \gamma^{-1}$. Again, in an ecological context, care must be taken with hyperprior choice; Figure 7(c) shows the marginal prior for $p_{ji}$ with a uniform prior on $\phi$ and an $\text{Exp}(1)$ prior on $\gamma$; panel (d) gives the equivalent plot under an $\text{Exp}(0.005)$ prior for $\gamma$, and shows more desirable behavior. An advantage of the beta second stage distribution is the ease in interpretation of probabilities. Independence of $p_{0i}$ and $p_{1i}$ would not seem reasonable a priori in many situations, however (for the Louisiana data the correlation between $\hat{p}_{00}, \hat{p}_{11}$ is 0.34), but it is not straightforward to incorporate dependence in the beta formulation. We now consider a prior that is statistically appealing and that allows dependence between the probabilities to be incorporated; the scale is unfortunately less easily interpreted, however.
Fig. 7. Marginal distribution of $p$ for various second stage and hyperprior specifications. Plots (a)–(d) consider $\text{Be}(a, b)$ second stage distributions and plots (e) and (f) normal distributions $N(\mu, \sigma^2)$ on the logits. Choices of hyperprior are: (a) $a$ and $b$ drawn from independent $\text{Exp}(0.5)$ distributions, (b) $a$ and $b$ drawn from independent $\text{Exp}(0.01)$ distributions, (c) uniform distribution on $a/(a+b)$, and independently $\text{Exp}(1)$ on $a+b$, (d) uniform distribution on $a/(a+b)$, and independently $\text{Exp}(0.005)$ on $a+b$, (e) logistic on $\mu$, and independently $\text{Ga}(0.1,0.1)$ on $\sigma^{-2}$, (f) logistic on $\mu$, and independently $\text{Ga}(1,0.01)$ on $\sigma^{-2}$. In all cases the marginal distribution for $p$ is collapsed over stages 2 and 3 of the hierarchy.
In this section we describe a flexible model that has not been previously used in the ecological context; we first describe the basic form, and then extensions. The model was suggested by Skene and Wakefield (1990) for multi-centre clinical trials with binary treatment and responses. In its most general form it provides a model that allows dependence by assuming the pair of logits arise from a bivariate normal distribution. Leonard (1972) describes an early approach to the analysis of a sequence of binomial experiments in which the logits were assumed to arise from a normal distribution. At the first stage of the hierarchy we take the convolution likelihood (or one of the approximations described in Section 5.6). Let \( \theta_{ji} = \log \left( \frac{p_{ji}}{1-p_{ji}} \right) \) denote the logit of probability \( p_{ji} \) in table \( i, i = 1, ..., m, j = 0, 1 \). Then at the second stage we assume

\[
\begin{align*}
\theta_{0i} &= \mu_0 + \delta_{0i} \\
\theta_{1i} &= \mu_1 + \delta_{1i}
\end{align*}
\]

(21)

where, initially, we assume that \( \delta_{0i} \) and \( \delta_{1i} \) are independent with \( \delta_{ji} \sim N(0, \sigma_j^2), j = 0, 1 \). To allow dependence we may assume that \( \delta_{0i}, \delta_{1i} \) arise from a zero mean bivariate normal distribution with variance-covariance matrix \( \Sigma \).

Area-level covariates may be incorporated in (21) by taking

\[
\begin{align*}
\theta_{0i} &= \mu_0 + \beta_0 z_i + \delta_{0i} \\
\theta_{1i} &= \mu_1 + \beta_1 z_i + \delta_{1i}
\end{align*}
\]

(22)

where \( z_i \) is an area-level covariate and \( e^{\beta_j} \) represents the ecological odds ratio corresponding to a unit increase in the area-level covariate \( z \). The normal distribution in (21) may also be replaced with the more robust t-distribution though this is likely to be of secondary importance for an overall population average (unless the number of tables is very small), though may provide improved inference for imputations in particular tables. Computation, which is described in Section 5.6, via Markov chain Monte Carlo (MCMC) is just as straightforward as in the normal case.

We now discuss hyperprior choice for the normal second stage models. As with the beta second stage, great care must be taken in hyperprior specification. Independent logistic distributions with location 0 and scale 1 for \( \mu_0, \mu_1 \) are on the surface appealing in the absence of substantive information since the distribution of the median \( p \) across the population of tables is then uniform. However, as we will expand upon, there is no “safe” choice when analysing ecological data, and a prior based on additional data or substantive knowledge is preferable. Many of the examples in WinBUGS User Manual (Spiegelhalter et al. 1998) use a Ga(0.001,0.001) prior for precisions of random effects. This choice is often not a good one (for an illustration in the context of spatial epidemiology see Kelsall and Wakefield, 1999), but in the ecological context it is very dangerous. The induced prior on an exchangeable \( p \) is shown in Figure 7(e) for a Ga(0.1,0.1) prior, the choice Ga(0.001,0.001) provides an even more extreme form with the mass concentrated close to 0 and 1. A more reasonable option is the choice of independent Ga(1,0.01) hyperpriors for \( \sigma_0^{-1}, \sigma_1^{-1} \). Figure 7(f) shows that the marginal distribution for \( p_{ji} \) that results is relatively uniform on (0,1). The marginal distribution for \( p \) with a second stage t \( _4 \) distribution (with logistic priors on the mean and equivalent gamma priors on the precisions) is virtually identical with that seen in Figure 7(f). Under the bivariate model an inverse Wishart prior may be placed on \( \Sigma \). This prior requires specification of a degrees of freedom parameter (with a proper prior requiring this value to be 2 or greater), and a scale matrix.

Corder and Wolbrecht (2004) have used the model described here to examine voting as a function of gender, in the period shortly after women became enfranchised in the United States in 1914. This application is difficult to answer from aggregate data alone because there is no possibility of collecting individual-level data, and the proportions of women is close to 50% in all areas. Corder and Wolbrecht (2004) incorporate both supplementary data and subjective opinion to carry out inference.
5.5. Spatial Epidemiological Hierarchical Models

We briefly describe a model that is related to that described in the previous section, for a spatial epidemiological study with a single binary exposure. Assuming a rare, non-infectious disease we assume

\[ Y_{ji} | \mu_j, \delta_i \sim \text{Poisson}(N_{ji} e^{\mu_j + \delta_i}) , \]

for \( j = 0, 1 \), and let \( \beta = \mu_1 - \mu_0 \) denote the log relative risk (which is assumed constant across areas). Upon aggregation we obtain

\[ Y_i | \mu_0, \beta, \delta_i \sim \text{Poisson}(N_i e^{\mu_0 + \delta_i} [(1 - x_i) + x_i e^\beta]) , \] (23)

\( i = 1, ..., m \). This model has the flavor of an ecological regression with the baseline probability varying by area. The random effects \( \delta_i \) may be assigned a spatial structure, or a combination of spatial and non-spatial random effects, as suggested by Besag et al. (1991) in a disease mapping context, and applied in geographical correlation studies by Clayton et al. (1993). An obvious choice for the spatial structure is the intrinsic conditional autoregressive model (also described in Besag et al. 1991) in which the mean of the conditional prior depends on the mean of the ‘neighbouring’ random effects, and the variance on the number of neighbours. A binomial version of the shared component model of Knorr-Held and Best (2001) would seem a plausible model for considering the spatial dependence between the logits of the two table-specific probabilities in order to allow a component of the spatial dependence to be common to both probabilities, but the lack of information will make inference from such models highly prior dependent unless there is supplementary data or additional information. When spatial models were investigated in this context by Haneuse and Wakefield (2004), inference was found to be highly dependent on prior modelling assumptions on the nature of the spatial dependence.

In geographical correlation studies the exposure of interest, \( z \) say, is often continuous. For ease of exposition we suppose that within area \( i \) the distribution of the exposures are independently distributed as \( z | \phi_i \sim p(\cdot | \phi_i) \) and for an individual in area \( i \) we assume the loglinear risk model \( \exp(\mu_0 + \beta z + \delta_i) \) where \( \delta_i \) may have spatial and non-spatial components. Under these circumstances the aggregate disease risk (which is analogous to \( q_i \)), is given by

\[ e^{\mu_0+\delta_i} \int e^{\beta z} p(z|\phi_i)dz. \]

As an example (Richardson et al. (1987)), if the within-area distribution of \( z \) were normal then \( \phi_i = (\bar{z}_i, s^2_i) \) where \( \bar{z}_i \) and \( s^2_i \) represent the mean and variance of the exposure, and we obtain an average risk of \( \exp(\mu_0 + \beta \bar{z}_i + \beta^2 s^2_i/2 + \delta_i) \). Wakefield and Salway (2001) consider such a parametric approach to ecological inference, and examine the effect of estimating the moments that are required. Plummer and Clayton (1996) have also considered this normal model.

As an alternative to the parametric approach, Prentice and Sheppard (1995) proposed an aggregate data approach in which within-area covariate samples \( z_{ij}, j = 1, ..., m_i \) are utilised to give aggregate risk

\[ \frac{e^{\mu_0+\delta_i} \sum_{j=1}^{m_i} e^{\beta z_{ij}}}{m_i} , \]

where \( \delta_i \) is an independent random effect. In the case of a binary \( z_i \) we obtain the marginal risk model

\[ e^{\mu_0+\delta_i} \{ (1 - \hat{x}_i) + \hat{x}_i e^\beta \} , \]

where \( \hat{x}_i \) is the estimate of the fraction in the area with \( z = 1 \). Hence we see by comparison with (23) that the approach is closely related to ecological regression with a rare outcome and an area-level random effect. Confounders may be included in both the parametric and aggregate data approaches though the characterisation of the multivariate distribution of exposures and confounders either via a probabilistic model, or sufficient samples, is likely to be difficult. This emphasizes the difficulty in controlling for ecological bias due to confounding.
In this section we will discuss computation for the Bayesian hierarchical models discussed in Sections 5.3 and 5.4. We have a generic three stage model with components

\[ Y_{ij} | p_{0i}, p_{1i}, p_{0i}, p_{1i}, \phi, \psi, \]

and \( \psi \) known. Computation for these models is naturally carried out via MCMC, a strategy that allows the conditional independencies in the hierarchy to be exploited. The major difficulty is with sampling from the conditional distribution for \( p_{0i}, p_{1i} \), which is given by

\[
\pi(p_{0i}, p_{1i} | N_{0i}, N_{1i}, y_i, \phi) \propto L(p_{0i}, p_{1i}) \times \pi(p_{0i}, p_{1i} | \phi),
\]

where the likelihood is given by the convolution (4). This distribution is analytically intractable, and for large marginal counts concentrates along the tomography line and is computationally expensive to evaluate. We first discuss an auxiliary variable scheme that is generally applicable, though computationally intensive for large marginal counts and non-rare events, before describing two approximations that are useful in particular situations. For each of the models discussed in Sections 5.3–5.5 it is straightforward to construct random walk chains for the conditional distribution of \( \phi \).

### Auxiliary Variable Scheme

With the introduction of auxiliary variables \( Y_{0i} \), one of the unobserved entries in Table 1, we have the (conditional) posterior

\[
\pi(y_{0i}, p_{0i}, p_{1i} | N_{0i}, N_{1i}, y_i, \phi) = \Pr(y_{0i} | N_{0i}, N_{1i}, y_i, p_{0i}, p_{1i}) \times \pi(p_{0i}, p_{1i} | N_{0i}, N_{1i}, y_i, \phi),
\]

where \( \pi(p_{0i}, p_{1i} | N_{0i}, N_{1i}, y_i, \phi) \) is given by (24). The second term of (25) is not of straightforward form (and is a posterior distribution of interest) and so sampling cannot be carried out directly. This formulation does suggest an MCMC scheme, however, alternating between the conditional distributions given by

\[
\Pr(y_{0i} | N_{0i}, N_{1i}, y_i, p_{0i}, p_{1i})
\]

(26)

and

\[
\pi(p_{0i}, p_{1i} | N_{0i}, N_{1i}, y_i, y_{0i}, \phi).
\]

Byers and Besag (2000) suggested this approach for extracting information from partially observed spatial epidemiological data with a rare outcome. In the example presented in their paper, disease mapping of male prostate cancer mortality in 3053 counties of the U.S. was carried out with adjustment for age. The association between prostate cancer and race was assumed constant across areas (analogously to ecological regression) and was estimated using the observed populations in the race by area margin, and by imputing counts for the race by mortality margin (which was unobserved).

The elegance of the above formulation is that the awkward bounds on \( Y_{0i}, Y_{1i} \) are automatically accommodated, and we avoid having to deal directly with the convolution distribution; the steps in the MCMC scheme are described in the Appendix. We here concentrate on the situation in which the table entries are large, since this is typically the scenario in social science applications. In this case the use of the auxiliary variable scheme is less appealing since: computation of (26), which is an extended hypergeometric distribution (see the Appendix) is extremely expensive; and the probabilities of the internal cells of the table concentrate on a relatively small number of values, producing slow movement around the space. Wakefield (2004) gives examples of these observations.

### Poisson Approximation

For a rare outcome (the situation considered by Byers and Besag, 2000), the binomial outcome in each row of Table 1 may be approximated by a Poisson distribution, in which case the convolution for the marginal
outcome $Y_i$ is a Poisson distribution also. Hence the likelihood is of the form

$$L(p_{0i}, p_{1i}) = \frac{\exp(-N_i q_i)(N_i q_i)^{y_i}}{y_i!},$$

where $q_i = N_i p_{0i} + N_{1i} p_{1i}$. One may work directly with this likelihood, or the introduction of the auxiliary variable $Y_0i$ again produces recognizable forms for the required conditional distributions.

**Normal Approximation**

With the normal approximation (17) it is straightforward to construct Metropolis-Hastings algorithms for $p_{0i}, p_{1i}$. This is the approach we take in Section 7, using the WinBUGS software (Spiegelhalter et al. 1998). For some data sets, and with the beta second stage model, we ran into difficulties using WinBUGS, because of sampling $p_{0i}, p_{1i}$ points very close to 0 and 1 (the likelihood is often non-negligible throughout the complete range). For these cases specialized code was written which implemented a random walk Metropolis-Hastings algorithm for the conditional distributions of $p_{0i}, p_{1i}$.

The use of the normal approximation alleviates the problem of computational expense of the likelihood evaluation, but convergence remains a problem with chains often displaying strong serial dependence and further research is required to provide algorithms that produce chains with improved properties.

5.7. **Consistency**

In the usual hierarchical modelling situation in which individual-level data are available, as more data are collected on each area, the hierarchical prior linking parameters across areas becomes less important for estimation of first stage parameters (so long as the prior does not assign zero mass to any region) since the information at the first stage dominates. In the limit, the likelihood tends to a point and inference will correspond to the examination of the likelihood for that table alone. Hence, in this situation, hierarchical models are, in a frequentist sense, consistent for the first stage parameters (at least as long as the likelihood is correctly specified).

Unfortunately in the ecological scenario the situation is dramatically different because as the data on each unit at the first stage (that is the margins for each table) increase, the fundamental difficulty of ecological inference remains. There is an intrinsic lack of information and any approach to point estimation that is based only on the marginal data gives a likelihood that is concentrated along a line (as illustrated in Figure 5(a)) and so the posterior will not tend to a point and consistency does not follow. As the number of tables $m$ increases, more information about the second stage distribution accrues, but the true values of $p_{0i}, p_{1i}$ will not be recovered since, informally speaking, the posterior tends to the product of a line and a density.

6. **Hazards of Hierarchical Modeling**

In this section we will illustrate that inference using hierarchical models can be highly sensitive to the choice of second and third stage prior distributions, even in the situation in which we have assumed the correct first and second stage models.

We simulate race-specific probabilities $p_{ji|a_j, b_j} \sim \text{Be}(a_j, b_j)$, where $a_j = 10, b_j = 30, j = 0, 1$ so that there is no difference between the races, and $\tilde{p}_0 = \tilde{p}_1 = 0.25$. We take $N_i = 10,000$ observations for each table. The number of tables is given by $m = n_x \times n_p^2$ where $n_x$ is the number of distinct values that $x$ takes, and $n_p$ is the number of distinct values taken by each of $p_0$ and $p_1$. The specific values are given by $x_k = c \times k / (n_x + 1), k = 1, \ldots, n_x$ with $0 \leq c \leq 1$ determining the distribution of the group proportions across areas; we take $c = 0.25, 0.375$ and 1, so that in the first two cases we have a minority group. The simulation is deterministic in the following sense. At the second stage take $p_{0i}$ and $p_{1i}$ to be the $k/(n_p + 1)$ quantiles, $k = 1, \ldots, n_p$, of
Table 5. Posterior means for simulated data. In all cases data were simulated as binomial random variables with \( N_i = 10,000 \) and \( p_i, a_j, b_j \sim \text{Be}(a_j, b_j) \), \( j = 0, 1 \), where \( a_0 = a_1 = 10, b_0 = b_1 = 30, \ i = 1, \ldots, m \), so that the true means, \( \bar{p}_0, \bar{p}_1 \) are 0.25 for both blacks and whites. At the third stage we have independent exponential priors with mean \( \lambda^{-1} \) for \( a_0, b_0, a_1, b_1 \).

<table>
<thead>
<tr>
<th>( m )</th>
<th>Design</th>
<th>( \lambda )</th>
<th>( \bar{p}_0 ) (95% interval)</th>
<th>( \bar{p}_1 ) (95% interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>( x \in (0, 0.25) )</td>
<td>0.5</td>
<td>0.27 (0.07,0.46)</td>
<td>0.25 (0.21,0.28)</td>
</tr>
<tr>
<td>27</td>
<td>( x \in (0, 0.25) )</td>
<td>0.5</td>
<td>0.34 (0.10,0.92)</td>
<td>0.22 (0.15,0.29)</td>
</tr>
<tr>
<td>27</td>
<td>( x \in (0, 0.375) )</td>
<td>0.5</td>
<td>0.25 (0.12,0.38)</td>
<td>0.25 (0.21,0.28)</td>
</tr>
<tr>
<td>27</td>
<td>( x \in (0, 0.375) )</td>
<td>0.5</td>
<td>0.24 (0.03,0.72)</td>
<td>0.24 (0.14,0.32)</td>
</tr>
<tr>
<td>27</td>
<td>( x \in (0, 1.00) )</td>
<td>0.5</td>
<td>0.25 (0.22,0.28)</td>
<td>0.25 (0.22,0.28)</td>
</tr>
<tr>
<td>27</td>
<td>( x \in (0, 1.00) )</td>
<td>0.5</td>
<td>0.26 (0.17,0.35)</td>
<td>0.26 (0.17,0.35)</td>
</tr>
<tr>
<td>27</td>
<td>Individual</td>
<td>0.01</td>
<td>0.25 (0.23,0.27)</td>
<td>0.25 (0.23,0.27)</td>
</tr>
<tr>
<td>27</td>
<td>Individual</td>
<td>0.01</td>
<td>0.26 (0.23,0.29)</td>
<td>0.26 (0.23,0.29)</td>
</tr>
<tr>
<td>64</td>
<td>( x \in (0, 0.25) )</td>
<td>0.5</td>
<td>0.25 (0.10,0.39)</td>
<td>0.25 (0.22,0.27)</td>
</tr>
<tr>
<td>64</td>
<td>( x \in (0, 0.25) )</td>
<td>0.5</td>
<td>0.28 (0.03,0.58)</td>
<td>0.25 (0.20,0.29)</td>
</tr>
<tr>
<td>64</td>
<td>( x \in (0, 0.375) )</td>
<td>0.5</td>
<td>0.25 (0.16,0.33)</td>
<td>0.25 (0.22,0.27)</td>
</tr>
<tr>
<td>64</td>
<td>( x \in (0, 0.375) )</td>
<td>0.5</td>
<td>0.24 (0.01,0.44)</td>
<td>0.25 (0.21,0.31)</td>
</tr>
<tr>
<td>64</td>
<td>( x \in (0, 1.00) )</td>
<td>0.01</td>
<td>0.25 (0.23,0.27)</td>
<td>0.25 (0.23,0.27)</td>
</tr>
<tr>
<td>64</td>
<td>( x \in (0, 1.00) )</td>
<td>0.5</td>
<td>0.25 (0.21,0.29)</td>
<td>0.25 (0.21,0.30)</td>
</tr>
<tr>
<td>64</td>
<td>Individual</td>
<td>0.01</td>
<td>0.25 (0.23,0.27)</td>
<td>0.25 (0.23,0.27)</td>
</tr>
<tr>
<td>64</td>
<td>Individual</td>
<td>0.01</td>
<td>0.25 (0.24,0.27)</td>
<td>0.25 (0.24,0.27)</td>
</tr>
<tr>
<td>125</td>
<td>( x \in (0, 0.25) )</td>
<td>0.01</td>
<td>0.24 (0.14,0.35)</td>
<td>0.25 (0.23,0.27)</td>
</tr>
<tr>
<td>125</td>
<td>( x \in (0, 0.25) )</td>
<td>0.01</td>
<td>0.25 (0.09,0.40)</td>
<td>0.25 (0.23,0.27)</td>
</tr>
<tr>
<td>125</td>
<td>( x \in (0, 0.375) )</td>
<td>0.01</td>
<td>0.25 (0.19,0.31)</td>
<td>0.25 (0.23,0.26)</td>
</tr>
<tr>
<td>125</td>
<td>( x \in (0, 0.375) )</td>
<td>0.01</td>
<td>0.25 (0.14,0.36)</td>
<td>0.25 (0.23,0.28)</td>
</tr>
<tr>
<td>125</td>
<td>( x \in (0, 1.00) )</td>
<td>0.01</td>
<td>0.25 (0.23,0.26)</td>
<td>0.25 (0.23,0.26)</td>
</tr>
<tr>
<td>125</td>
<td>( x \in (0, 1.00) )</td>
<td>0.01</td>
<td>0.25 (0.23,0.27)</td>
<td>0.25 (0.23,0.28)</td>
</tr>
<tr>
<td>125</td>
<td>Individual</td>
<td>0.01</td>
<td>0.25 (0.24,0.26)</td>
<td>0.25 (0.24,0.26)</td>
</tr>
<tr>
<td>125</td>
<td>Individual</td>
<td>0.01</td>
<td>0.25 (0.24,0.27)</td>
<td>0.25 (0.24,0.27)</td>
</tr>
</tbody>
</table>

For the more asymmetric \( x \) distributions in particular, the posterior distributions for \( p_{0i} \) are very flat, reflecting the lack of information, and the Markov chains were extremely slow in their mixing; up to ten million iterations were used for each analysis.

The simulated data were analysed using a hierarchical model with the normal approximation to the convolution at the first stage, independent beta priors at the second stage, and exponential hyperpriors with mean \( \lambda^{-1} \) at the third stage. We chose two values for the latter, \( \lambda = 0.5, 0.01 \), the former value was used by King et al. (1999) and the latter was selected for comparison (the marginal distributions for \( p \) under these priors are shown in Figure 7(a) and (b)).

Table 5 shows posterior summaries for each of the 18 simulations corresponding to \( \lambda = 0.01, 0.5, m = 27, 64, 125, c = 0.25, 0.375, 1.0 \). Individual data from \( m \) binomial distributions, again with \( N_i = 10,000 \) individuals per area are also included for comparison purposes. Figure 8 shows the tomography lines reflecting the bounds on \( p_0, p_1, i = 1, \ldots, m \) for each of the \( c, m \) combinations.

In all cases the true values of \( a_0 = a_1 = 10 \) and \( b_0 = b_1 = 30 \) are bracketed between the posterior means under the two hyperpriors (which have expectations of 2 and 100). The individual rows allow a calibration by showing the accuracy of estimation when individual-level data are available.

A striking aspect of the simulations is the sensitivity of inference for \( \bar{p}_0, a_0, b_0 \) to the hyperprior specification.
Fig. 8. Tomography lines and true \( p_{0i}, p_{1i}, i = 1, \ldots, m \), for: (a) \( m = 27, c = 0.25 \), (b) \( m = 27, c = 0.375 \), (c) \( m = 27, c = 1 \), (d) \( m = 64, c = 0.25 \), (e) \( m = 64, c = 0.375 \), (f) \( m = 64, c = 1 \), (g) \( m = 125, c = 0.25 \), (h) \( m = 125, c = 0.375 \), (i) \( m = 125, c = 1 \).
Examination of the non-ecological simulations in which the individual-level data are available reveals that although the estimation of \(a\) and \(b\) is sensitive to the hyperprior choice, the estimation of \(\hat{p}\) is not. The extent of the sensitivity in the ecological simulations is a function of the range of \(x\) (as measured through \(c\)), and the sample size \(m\). The first column of Figure 8 displays the tomography lines for \(c = 0.25\) and \(m = 27, 64, 125\) and illustrates the lack of information that is available for \(p_{0i}\). When \(c = 0.25, 0.375\) the information to estimate \(p_{0i}\) is not arising from the likelihood, but from the beta-exponential prior, and from the value of \(p_{1i}\) (due to the strong dependence between these quantities in the likelihood). The tomography lines are relatively flat for the \(c = 0.25\) cases and so there is extreme sensitivity to the exact form of these two sources of information, which explains why care is required in prior choice (at both stages two and three). Again, this emphasizes the need for a “good” design, that is, one which has some areas with \(x_i\) close to 0, and some with \(x_i\) close to 1.

For the \(c = 0.25, m = 27\) case there is serious inaccuracy in \(\hat{p}_0\). From Table 4 we know that the MLEs are \(\hat{p}_{0i} = 1\) in this scenario which would apparently explain the observed bias. However, when the same data were analyzed with the logistic normal model of Section 5.4, with \(\text{Ga}(0.0001, 0.0001)\) hyperpriors for the precisions, no bias results, suggesting that for this case the observed bias could be due to numerical instability arising from the hyperprior placing too much weight on small values of \(a\) and \(b\). The truncated normal model also gave accurate point estimates of \(\hat{p}_0\) for each of the simulated data sets, and the logistic normal model showed much less sensitivity to the hyperprior choice.

We have only presented very limited simulations with a specific second and third stage combination here, but it is clear that one should be very careful with prior specification at both stages, particularly when there is asymmetry in the distribution of the minority fraction \(x_i\) across areas. In a real data analysis the sensitivity of inference to this choice must be examined, as we illustrate in the next section.

7. Hierarchical Modelling of the Louisiana Data

Apart from the truncated normal model each of the hierarchical models described in this section were analysed using the WinBUGS software, Spiegelhalter et al. (1998), using the normal approximation to the convolution likelihood. Convergence was assessed informally via examination of trace plots and quantiles across multiple chains. Due to the lack of information in the data the chains were often slowly mixing, and chains were run for up to two million iterations. Convergence was also dependent on the hyperprior specification.

In general we recommend a strategy in which a range of second and third stage models are fitted to examine the sensitivity of inference; assessing one choice over another is likely to be extremely difficult with the ecological data alone, however. 

**Truncated Normal Model**

The truncated normal model was fitted using the EzI software, Benoit and King (1998). Figure 1 shows that many of the tomography lines are flat for \(\hat{p}_0\), and cover \((0,1)\). Table 6 shows that inference for the whites is relatively accurate but for the blacks is poor.

**Beta Model**

As in Section 6 we use, for illustration, two hyperprior specifications. The \(\text{Exp}(0.5)\) hyperprior is not consistent with the data and produced poorer mixing than the \(\text{Exp}(0.01)\) choice.

In Table 6 we show that the seemingly innocuous choice of King et al. (1999) of exponential hyperpriors with mean 2 can be a poor choice in some situations. Using a simple method of moments estimator and the individual-level data, we obtain estimates for \(\hat{a}_0 = 4.5, \hat{b}_0 = 142, \hat{a}_1 = 3.5, \hat{b}_1 = 15\) (these values are close to those obtained when analyzing the individual-level data under a hierarchical model), so that \(b_0\) in particular is large. For an \(\text{Exp}(0.5)\) random variable the probabilities of exceedance of 5 and 10 are 0.082 and 0.007, respectively while for an \(\text{Exp}(0.01)\) random variable these probabilities are 0.95 and 0.90, so that far larger values of the hyperparameters are supported by the latter prior.
Table 6. Bias in population-averaged estimated fractions of Republican party registration for the Louisiana data for race \( j \), for a variety of hierarchical models; the % bias is given by \( 100 \times (\hat{p}_j - p_j) / p_j, j = 0, 1 \), with \( \hat{p}_j \) given by (9). The posterior medians are reported. True (observed) population-weighted fractions are 0.035 and 0.25.

<table>
<thead>
<tr>
<th>Second Stage</th>
<th>Hyperpriors</th>
<th>( \hat{p}_0 ) (% bias)</th>
<th>( \hat{p}_1 ) (% bias)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truncated normal</td>
<td>–</td>
<td>0.11 (203)</td>
<td>0.23 (-11)</td>
</tr>
<tr>
<td>Independent Beta, Be((a_j, b_j))</td>
<td>Independent Exp(0.01)</td>
<td>0.072 (106)</td>
<td>0.24 (-5.6)</td>
</tr>
<tr>
<td>Independent Beta, Be((a_j, b_j))</td>
<td>Independent Exp(0.5)</td>
<td>0.11 (210)</td>
<td>0.20 (-19)</td>
</tr>
<tr>
<td>Independent Beta, Be((a_j, b_j))</td>
<td>Uniform on (a_j/(a_j + b_j)), Exp(0.005) on (a_j + b_j)</td>
<td>0.070 (101)</td>
<td>0.24 (-5.3)</td>
</tr>
<tr>
<td>Independent Normal, (N(\mu_j, \sigma_j^2))</td>
<td>Logistic on (\mu_j), Ga(0.001,0.001) on (\sigma_j^{-2})</td>
<td>0.062 (77)</td>
<td>0.24 (-4.1)</td>
</tr>
<tr>
<td>Independent Normal, (N(\mu_j, \sigma_j^2))</td>
<td>Logistic on (\mu_j), Ga(1.0,0.01) on (\sigma_j^{-2})</td>
<td>0.047 (36)</td>
<td>0.25 (-1.9)</td>
</tr>
<tr>
<td>Independent Student (t_4(\mu_j, \sigma_j^2))</td>
<td>Logistic on (\mu_j), Ga(1.0,0.005) on (\sigma_j^{-2})</td>
<td>0.075 (117)</td>
<td>0.24 (-6.1)</td>
</tr>
<tr>
<td>Bivariate normal (N(\mu, \Sigma))</td>
<td>Logistic on (\mu_j), Wishart for (\Sigma^{-1}) (correlation=0)</td>
<td>0.061 (77)</td>
<td>0.24 (-4.0)</td>
</tr>
<tr>
<td>Bivariate normal (N(\mu, \Sigma))</td>
<td>Logistic on (\mu_j), Wishart for (\Sigma^{-1}) (correlation=0.5)</td>
<td>0.046 (31)</td>
<td>0.25 (-1.7)</td>
</tr>
<tr>
<td>Independent normal, (N(\mu_j, \sigma_j^2))</td>
<td>Logistic on (\mu_j), Ga(1.0,0.01) on (\sigma_j^{-2})</td>
<td>0.051 (48)</td>
<td>0.25 (-2.5)</td>
</tr>
<tr>
<td>Contextual, black/white</td>
<td>Independent (N(0, 0.35^2)) on (\beta_0, \beta_1)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Independent normals, (N(\mu_j, \sigma_j^2))</td>
<td>Logistic on (\mu_j), Ga(1.0,0.01) on (\sigma_j^{-2})</td>
<td>0.046 (33)</td>
<td>0.25 (-1.8)</td>
</tr>
<tr>
<td>Contextual, white only</td>
<td>(N(0, 0.35^2)) on (\beta_1)</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

For the Louisiana data the estimate of \( \hat{p}_0 \) under the Exp(0.5) prior is 50% larger that for the Exp(0.01) prior because \( b_0 \) is underestimated by the former (posterior mean 6.6) and overestimated slightly by the latter (posterior mean 168). As in Section 6 we have bias towards 0.5, reflecting the distribution of \( x \) values across areas and flat tomography lines covering (0,1) for many areas (Figure 1). Replacing the normal approximation to the convolution by the binomial at the first stage (with Exp(0.01) hyperpriors) produced little change in the population average fractions, due to the large values of \( N_i \) in this example. The results from the analysis with a uniform prior on \( a/(a+b) \) and Exp(0.005) on \( a+b \) are very similar to the Exp(0.01) priors case.

**Normal Model**

Considering first the independent normal specifications, the Ga(0.001,0.001) prior produced similar estimates to the Exp(0.01) beta analysis, but the estimates under the more reasonable Ga(1.0,0.01) prior (Figure 7(f)) were superior. These models do not allow contextual effects or dependence between the logits, which goes some way towards explaining the discrepancy with the true summaries. Independent Student \( t_4 \) distributions produced less accurate estimates than the normal alternative for this example.

A bivariate normal distribution was fitted with the prior for the precision matrix for the logits of \( p_{0i}, p_{1i} \) being a Wishart distribution with 2 degrees of freedom, and precision matrix with diagonal elements of 0.02, and an off-diagonal element of zero. This prior has the same marginal mean and variance as the Ga(1.0,0.01) prior, and zero prior correlation between the logits of black and white probabilities across areas. The resultant Markov chain was highly unstable, suggesting that there was insufficient information in the data to estimate the additional correlation parameter. The posterior median (95% credible interval) for the correlation between the logits was 0.89 (-0.97,0.99) reflecting this uncertainty. Accurate estimates were obtained (Table 6) but the instability in the chain gives reason to question the accuracy of posterior summaries. Also reported are summaries from an analysis in which a Wishart prior was taken with 2 degrees of freedom, diagonal elements 0.027 and off-diagonal element -0.013. This produces a prior with the same marginal mean and variance as the Ga(1.0,0.01) prior and, roughly-speaking, has a correlation of 0.5 between the logits of black and white probabilities across areas. The posterior median (95% credible interval) for the correlation between the logits was 0.74 (-0.99,0.98) again reflecting high uncertainty.

Finally the model with independent normal random effects and contextual effects (22) with \( z_i = x_i \) was fitted with logistic priors on \( \mu_0 \) and \( \mu_1 \). For the regression coefficients \( \beta_0 \) and \( \beta_1 \) we assume \( N(0, 0.35^2) \) priors. These reflect the belief that with probability 0.95 the ecological odds ratio associated with the contextual effects will be between 1/2 and 2 (giving a standard deviation of \( \log 2/1.96 = 0.35 \)). The resultant Markov chain was very slowly mixing but produced relatively accurate estimates. The posterior medians (95% credible intervals) are -0.03 (-0.72,0.66) for \( \beta_0 \) (the contextual effect for blacks) and 0.67 (-0.55,0.69) for \( \beta_1 \).
The extension of the hierarchical models described in Section 5 to the $R \times C$ case has been considered by Brown and Payne (1986) and Rosen et al. (2001). The notation we adopt is summarized in Table 7 with the individual-level data of each table is the product of the individual-level data by the contextual effect due to race. Unfortunately this example illustrates that the likelihood alone does not provide sufficient information for simultaneous estimation of all of these parameters, unless substantive prior information is available (or additional data).

This example was chosen because the internal cells were available, thus allowing a comparison of the methods that are reviewed and proposed here. For these data, hierarchical models have performed reasonably well, but would not have been so successful if strong contextual effects had been present (without prior information on their size being available). In general, a priori we would believe in both positive dependence between contextual effects due to race. Unfortunately this example illustrates that the likelihood alone does not provide sufficient information for simultaneous estimation of all of these parameters, unless substantive prior information is available (or additional data).

8. $R \times C$ Tables

The extension of the hierarchical models described in Section 5 to the $R \times C$ case is relatively straightforward; this case has been considered by Brown and Payne (1986) and Rosen et al. (2001). The notation we adopt is summarized in Table 7 with $R$ being the number of levels for the predictor variable, and $C - 1$ the number of distinct response categories. Analogously to the $2 \times 2$ case, a plausible sampling scheme for the individual-level data of each table is the product of $R$ multinomial distributions $M_C(N_{ri}, p_{ri})$, where $p_{ri} = (p_{r1}, ..., p_{rC})$ and $E[Y_{ric}|p_{ric}] = N_{ri} \times p_{ric}$, $r = 1, ..., R$.

The distribution of the aggregate counts $Y_{i1}, ..., Y_{i,C-1,i}$ is a convolution of multinomial distributions, but if the counts are large we can approximate the distribution by the $C - 1$ dimensional normal distribution with means $\sum_{r=1}^{R} N_{ri}p_{ric}$, variances $\sum_{r=1}^{R} N_{ri}p_{ric}(1 - p_{ric})$, $c = 1, ..., C - 1$, and covariances $-\sum_{r=1}^{R} N_{ri}p_{ric}p_{ric'}$, $c \neq c'$, $c, c' = 1, ..., C - 1$. These forms have been used by a number of authors including Hawkes (1969) and Brown and Payne (1986). Rosen et al. (2001) use a multinomial likelihood with marginal mean equal to the mean of the convolution but with a variance which is smaller.

The extension of hierarchical approaches is straightforward. The beta model is generalized to the Dirichlet, and for the logit model we can take $\theta_{ric} = \log \left( \frac{p_{ric}}{1 - p_{ric}} \right)$, to be the logit, relative to the final category (say), $r = 1, ..., R$, $c = 1, ..., C - 1$. Letting $\theta_{ri} = (\theta_{r1}, ..., \theta_{r_C})$, a simple model assumes that the collection $\theta_{1i}, ..., \theta_{ri}$ arise from a $C - 1$ dimensional normal distribution. MCMC algorithms, along the lines of those described in Section 5.6 may again be constructed.

9. Combining aggregate and individual level data

We now consider the situation in which survey data are available, Table 8 illustrates the notation in this case, the observed counts are $z_{0i}, z_{1i}$ and $y_{i}, i = 1, ..., m$. When such survey data are available on a subset of individuals within particular areas then the resultant product of binomial distributions may be simply

<table>
<thead>
<tr>
<th>$Y = 1$</th>
<th>$Y = 2$</th>
<th>...</th>
<th>$Y = C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = 1$</td>
<td>$Y_{11i}$</td>
<td>...</td>
<td>$Y_{1Ci}</td>
</tr>
<tr>
<td>$x = 2$</td>
<td>$Y_{21i}$</td>
<td>...</td>
<td>$Y_{2Ci}</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$x = R$</td>
<td>$Y_{RIi}$</td>
<td>...</td>
<td>$Y_{RCi}</td>
</tr>
</tbody>
</table>

| $Y_{i1}$ | $Y_{i2}$ | ... | $Y_{iC}$ | $N_{i}$ |

(table 7)

Jon Wakefield
We illustrate the general approach using the Louisiana registration/race data. We consider a hierarchical model in which the binomial and normal approximation to the convolution likelihoods are combined, and the race-specific probabilities are assumed to arise from independent second-stage beta distributions (in additional analyses the latter choice was seen to be unimportant). To mimic the situation in which survey data are available to supplement aggregate data, we first sample 5% of observations from within each table; so a number of situations in which different data are available and Table 9 gives numerical summaries in terms of the population race-specific registration fractions in these situations.

Figure 9 plots the estimated black and white fractions of Republican registration versus the true values, for the normal second stage model, the black probabilities are generally overestimated, with corresponding slight underestimation in the white probabilities. Panels (c) and (d) show the equivalent plots under the beta model with the same, though more pronounced, direction of bias. Panels (e) and (f) summarize the situation in which the survey data only are available. Unbiased estimation results, but there is large variability about the line of equality, particularly for the blacks.

When the aggregate and 5% survey data are combined, the variability is reduced, showing that aggregate data do provide valuable additional information (panels (g) and (h)). Finally we consider the situation when the aggregate data is supplemented by survey data from one race only. When the white data from the 5% sample only are added (panels (i) and (j)), some bias remains, but the situation is vastly improved over the aggregate data only case. The most interesting result occurs when the black individual data are added. Even though these data are far less numerous than the white data only (1.4% are black and 3.6% white), the results (panels (k) and (l)) are superior to the white subsample analysis. No convergence difficulties were encountered showing that supplementing the aggregate data with just a small amount of survey data leads to well-behaved posteriors. These empirical results are intriguing since they indicate that the ecological inference problem can be avoided with relatively small, carefully-chosen, samples and illustrates that aggregate data are beneficial in that they can add precision to individual-level estimates.

In epidemiology, outcome-based sampling designs have a long history with the most well-known example being the case-control study (e.g. Breslow and Day, 1980); in economics, such sampling is known as choice-based sampling (Manski and McFadden, 1981). The advantages of these designs is that they maximise the information concerning the smallest cell of the $2 \times 2$ table, hence increasing precision. The minority population are playing the role of the cases in this setting. Another important question is in which areas these data should be collected, and of what sizes the samples should be. The literature on so-called two-phase designs may be useful in this regard. In such designs the initial sample (for which limited information only is available) is supplemented with a second sample, where sampling is based on the observed responses from the

<table>
<thead>
<tr>
<th>$x = 0$</th>
<th>$x = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y = 0$</td>
<td>$z_{0i}$</td>
</tr>
<tr>
<td>$Y = 1$</td>
<td>$z_{1i}$</td>
</tr>
<tr>
<td>$m_i - z_i$</td>
<td>$z_i$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$Y = 0$</th>
<th>$Y = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{0i} - m_{0i}$</td>
<td></td>
</tr>
<tr>
<td>$N_{1i} - m_{1i}$</td>
<td></td>
</tr>
</tbody>
</table>
Fig. 9. Estimated Republican registration probabilities versus the truth (observed fractions). The estimates are posterior medians based upon a hierarchical model, see Table 9 and the text for details of the model specification. Aggregate data only, normal model (a) blacks, (b) whites; Aggregate data only, beta model (c) blacks, (d) whites; Individual 5% sample only: (e) blacks, (f) whites; Aggregate and individual 5% data (g) blacks, (h) whites; Aggregate and individual black data; (i) blacks, (j) whites; Aggregate and individual white data; (k) blacks, (l) whites.

Table 9. Bias in population-averaged probabilities of voter Republicans registration, $p_j = \sum N_j \hat{p}_{j_i} / \sum N_j$, $j = 0, 1$, for various aggregate and survey data scenarios. Unless otherwise stated, the second stage distribution is a product of beta distributions with Exp(0.01) hyperpriors. The summary measures in columns 4 and 5 are given by $S_j = \sum |\hat{p}_{j_i} - \hat{p}_{j_i}| / \hat{p}_{j_i}$, $j = 0, 1$, and represent a relative measure of the deviations of the area-race specific estimated fractions from $\hat{p}_{j_i}$, the true fractions.

<table>
<thead>
<tr>
<th>Data Source</th>
<th>$\hat{p}_0$ (% bias)</th>
<th>$\hat{p}_1$ (% bias)</th>
<th>$S_0$</th>
<th>$S_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All individual</td>
<td>0.035 (0)</td>
<td>0.25 (0)</td>
<td>1.8</td>
<td>0.1</td>
</tr>
<tr>
<td>Aggregate beta second stage</td>
<td>0.068 (96)</td>
<td>0.24 (-5.0)</td>
<td>107</td>
<td>6.2</td>
</tr>
<tr>
<td>Aggregate normal second stage</td>
<td>0.048 (39)</td>
<td>0.24 (-2.1)</td>
<td>73</td>
<td>4.2</td>
</tr>
<tr>
<td>5% individual sample</td>
<td>0.034 (-3.5)</td>
<td>0.25 (-0.039)</td>
<td>17</td>
<td>4.0</td>
</tr>
<tr>
<td>Aggregate and 5% combined sample</td>
<td>0.034 (-2.9)</td>
<td>0.25 (0.16)</td>
<td>16</td>
<td>0.9</td>
</tr>
<tr>
<td>Aggregate and 1.6% black sample</td>
<td>0.034 (-3.5)</td>
<td>0.27 (4.7)</td>
<td>17</td>
<td>2.4</td>
</tr>
<tr>
<td>Aggregate and 3.4% white sample</td>
<td>0.033 (-6.1)</td>
<td>0.25 (0.32)</td>
<td>29</td>
<td>1.6</td>
</tr>
</tbody>
</table>
first sampling (in the ecological context the first phase could represent the aggregate data). There is a large literature on such designs, see for example White (1982), Breslow and Cain (1988), Breslow and Holubkov (1997), Scott and Wild (1997), Whittemore (1997) and Breslow and Chatterjee (1999). As pointed out by Breslow and Cain (1988), efficiency can be increased by not only using both stages of data (including the first phase marginal data for which covariate data are unavailable), but also by choosing the second-stage sample judiciously.

Conventional survey sampling designs (e.g. Korn and Graubard, 1999) consider multistage sampling where sampling is carried out on the basis of stratification variables (e.g. stratified simple random sampling), or for convenience by selecting geographical areas and sampling within these. Again the benefits of judicial sampling are exploited.

In the aggregate situation that we are considering, we observe marginal response and ‘exposure’ (i.e. black/white) variables, with stratification by area. This differs from the two-phase approach in which data cross-classified by response and exposure are available, and multistage approaches in which stratification variables only are available, but the links are apparent. The design problem of the fractions to sample within each area will depend on both \( q_i \) and \( x_i, i = 1, ..., m \).

Once individual data are available then model checking becomes possible. In particular, one may distinguish between ecological regression and linear neighbourhood models. Contextual effects may also be more reliably estimated. If contextual effects are present but not modelled then imputations for individual areas may still be accurate, since they are predictions but modelling is required to examine a causal relationship between, for example, the racial composition of a neighbourhood and registration.

An important final point is that the individual-level (survey) data must be representative and accurate which will be difficult in some situations. For example, in political science applications voters do not truthfully answer questions on whom they voted for. For example, Burden (2000) discusses the non-representative nature of surveys of voter turnout, and in particular the problem of over-reporting. If there is a suspicion of bias, then sensitivity to this aspect can be addressed also.

10. Discussion

In this paper we have considered inference for \( 2 \times 2 \) tables in which the margins only are observed. By consideration of a convolution likelihood the fundamental indeterminacy of ecological inference has been clarified. In particular the importance of the distribution of the minority group across areas has been illustrated. Hierarchical modeling using the convolution likelihood provides a plausible model in a number of situations. However, without external information, either supplementary data or prior knowledge, one cannot exclude contextual explanations for the observed behaviour of marginal fractions responding. The sensitivity to hyperprior specification has been illustrated indicating that great care should be taken in choosing hyperpriors and is preferably carried out using substantive prior information. If such information is unavailable then sensitivity analyses should be used, for example to investigate plausible strengths of contextual effects. A more in-depth examination of the effect of hyperpriors on different second stage models is also required. The use of the Bayesian hierarchical models described in Section 5 in real data situations such as that described by Corder and Wolbrecht (2004) would be highly desirable to refine the methodology.

A number of issues require further investigation. A closer examination of the accuracy of the normal approximation as a function of \( N_i \) is required. We have not discussed diagnostics for assessing the appropriateness of distributional choices as we believe the lack of information in ecological data makes this extremely difficult though clarification of this issue would be welcome. It would appear desirable to control for as much variability in the response as possible, through the introduction of additional predictors (as is common in epidemiology), though the joint distribution of all predictors within an area is desirable, and this is not likely to be routinely available. The computational schemes that we have utilized are relatively unsophisticated and there is a need for more efficient algorithms.

We end with the obvious statement that the solution to the ecological inference problem is to supplement the
aggregate data with representative and accurate survey sample information on individuals within the areas of interest, and/or use prior information concerning the within-area probabilities and the extent of contextual effects. Each of these possibilities may be formally considered within the hierarchical models described in Section 5. The discussion of Section 9 indicates that dramatic improvements in inference may be obtained with small carefully-chosen samples. Future work will exploit existing methods in the epidemiological and survey sampling literature to produce guidelines for the sampling of individual data, when such sampling is feasible.

Appendix

For the MCMC scheme outlined in Section 5.6 we discuss the sampling of the auxiliary variable \( y_{0i} \) and the probabilities \( p_{0i}, p_{1i} \).

**Auxiliary Variable Sampling:**

For the missing data \( y_{0i} \), the distribution (26) is an extended hypergeometric distribution (e.g. Johnson and Kotz, 1969) with margins \( N_{0i}, N_{1i}, y_i, N_i - y_i, i = 1, \ldots, m \):

\[
\Pr(Y_{0i} = y_{0i}|\theta_i, N_{0i}, N_{1i}, y_i) = \begin{cases} \frac{\binom{N_{0i}}{y_{0i}} \binom{N_{1i}}{y_i - y_{0i}} \theta_i^{y_{0i}}}{\sum_{u = i}^{u_u} \binom{N_{0i}}{u} \binom{N_{1i}}{y_i - u} \theta_i^u} & y_{0i} = l_i, \ldots, u_i, \\ \text{otherwise} & \end{cases}
\]

(28)

where \( l_i = \max(0, Y - N_{1i}) \) and \( u_i = \min(N_{0i}, Y) \) and \( \theta = p_{0i}(1 - p_{1i})/p_{1i}(1 - p_{0i}) \) is the odds ratio in the table. This discrete distribution may be sampled from in an obvious fashion, but in typical political science/sociology applications the margins are large and so generation is highly inefficient due to the summation over a large number of terms, each of which contains factorials. This problem has been considered (for this and more general conditional inference problems) by, amongst others, Forster, McDonald and Smith (1996), Strawderman and Wells (1998), Booth and Butler (1999), Mehta, Patel and Senchaudhuri (2000) and Liao and Rosen (2001). Byers and Besag (2000) considered a rare outcome, in which case the extended hypergeometric random variables may be approximated by Poisson random variables. The moments of the extended hypergeometric distribution are numerically difficult to calculate for tables with large margins and a number of approaches to their calculation have been suggested (McCullagh and Nelder, 1989, p. 259; Satten and Kupper, 1990; Liao, 1992, Liao and Rosen, 2001).

The mode is available in closed form, however, and the distribution is log-concave, which suggested the following approach. Starting from the mode we move outwards until a “small” value on each side is found, at these points the distribution is truncated (to make the summation over fewer terms) and the whole distribution is normalized. To decide upon the cut-off we note that for a normal distribution we can evaluate the ratio of the density at the mode to the density at a specified quantile, and so obtain a “small” value (assuming that the distribution is approximately normal in shape). This approach has the advantage of sorting the probabilities so the inversion method is more efficient. Care must be taken to evaluate the unnormalised distribution on the log scale, and scale via the largest unnormalised value, so that underflow/overflow does not occur. We have successfully implemented this scheme for problems with small to medium sized marginal tables. For large marginal tables the approach is less appealing due to slow mixing; Wakefield (2004) presents an example of the poor convergence behaviour of the algorithm for a large table.

**Posterior Probability Sampling:**

Here we are required to sample from the conditional distribution (27), for \( i = 1, \ldots, m \). If we assume \( p_{ji}|a_j, b_j \sim \text{Be}(a_j, b_j) \), then this conditional distribution corresponds to the product

\[
\text{Be}(y_{0i} + a_0, N_{0i} - y_{0i} + b_0)\text{Be}(y_{1i} + a_1, N_{1i} - y_{1i} + b_1),
\]

(29)
for $i = 1, ..., m$, and is straightforward to sample from. With a normal second stage distribution for the logits, the conditional distribution is no longer of standard form but a Metropolis-Hastings step is easy.

For a frequentist approach an EM algorithm would be simple to construct when one considers the auxiliary variables as missing data, see Little and Rubin (1987) for a general description and McCullagh and Nelder (1989, p. 338) for a discussion in the context considered here. Maximization of the extended hypergeometric full likelihood is feasible with the algorithms for obtaining the mean cited above, though convergence may be poor. Brown and Payne (1986) suggest such an approach for their ACM model but report prohibitively slow convergence.

In many political science applications (in particular) the counts $Y_{0i}, Y_{1i}$, or equivalently the fractions $Y_{0i}/N_{0i}$, $Y_{1i}/N_{1i}$, are of specific interest. From a Bayesian perspective we may treat these counts as random variables and then sample from

$$
\Pr(Y_{0i}, Y_{1i}|N_{0i}, N_{1i}, y) = \int \Pr(Y_{0i}, Y_{1i}|N_{0i}, N_{1i}, y_{i}, p_{0i}, p_{1i})\pi(p_{0i}, p_{1i}|N_{0i}, N_{1i}, y_{i})dp_{0i}dp_{1i}. \quad (30)
$$

If we can simulate from $\Pr(Y_{0i}, Y_{1i}|N_{0i}, N_{1i}, y_{i}, p_{0i}, p_{1i})$ then it is straightforward to simulate from (30), once samples for $p_{0i}, p_{1i}$ are available from $\pi(p_{0i}, p_{1i}|N_{0i}, N_{1i}, y_{i})$. This is equivalent to the distribution of $y_{0i}$ given the row and column margins, and the table probabilities, which is the extended hypergeometric (28). Hence (30) is an overdispersed extended hypergeometric distribution. We obtain the second unobserved cell as $Y_{1i} = Y_{i} - Y_{0i}$ and the resultant fractions satisfy the tomography identity.

Acknowledgments

I would like to thank Mark Handcock, Patrick Heagerty, Thomas Lumley, Ruth Salway, Dave Stephens and Jon Wellner for useful discussions. The comments of the editor, screener, and four referees greatly aided in the structuring of the paper. I would also like to thank Julian Besag for pointing me in the direction of Byers and Besag (2000), Gary King and Steve Voss for supplying data in a convenient form, and Sebastien Haneuse for assistance with the computation.

References


Rosen, O., Jiang, W., King, G. and Tanner, M.A. (2001). Bayesian and frequentist inference for ecological


