Ecological Inference in the Social Sciences

Adam Glynn[1] and Jon Wakefield[2]
[1]Department of Government, Harvard University, Cambridge.
[2]Departments of Statistics and Biostatistics, University of Washington, Seattle.

May 17, 2009

Abstract

Ecological inference is a problem of partial identification, and therefore precise conclusions are rarely possible without the collection of individual level (identifying) data. Without such data, sensitivity analyses provide the only recourse. In this paper we review and critique recent approaches to ecological inference in the social sciences, and describe in detail hierarchical models, which allow both sensitivity analysis and the incorporation of individual level data into an ecological analysis. A crucial element of a sensitivity analysis in such models is prior specification, and we detail how this may be carried out. Furthermore, we demonstrate how the inclusion of a small amount of individual level data from a small number of ecological areas can dramatically improve the properties of such estimates.

1 Introduction

In this paper the problems of making individual-level inference from ecological data is considered. In particular suppose we have a set of $R \times C$ tables in which only the margins 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; Wakefield, 2008); King (1997) and Cleave et al. (1995) describe further application areas. However, the exact nature of the inferential question differs importantly across these disciplines (Salway and Wakefield, 2004).

In epidemiological applications the usual aim is to estimate a risk contrast between exposed and unexposed individuals in a certain population and time period. This contrast can then be used to predict the number of new cases in a future time period (for public health provision, for example) or to make causal inferences. 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, race, 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 2×2 tables, since within each area there will (typically) be multiple confounders for which to control. 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., 2000; Wakefield, 2008).

By contrast, in the social sciences, ecological inference can have causal and non-causal inferential purposes. In many cases, social scientists have concentrated on imputing the missing cells in the constituent 2×2 or $R \times C$ tables. This type of ecological inference is often referred to as EI (named after the software package that accompanies King's 1997 book). For example, in political science, a 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. However, in many recent cases, EI is followed by a second stage analysis that utilizes the imputed data, often as the dependent variable in a regression (Herron and Shotts, 2003b) and may be implicitly causal. For example, Burden and Kimball (1998) used EI to analyze ticket splitting rates across congressional districts, and then used the estimated rates as a dependent variable in a second stage analysis to determine why voters were splitting their tickets. This type of analysis is often referred to as EI-R.

Although the shortcomings of ecological inference in the EI and epidemiological contexts have been documented (Achen and Shively, 1995; Greenland and Robins, 1994; Cho, 1998; Freedman et al., 1998; Gelman et al., 2001; Wakefield, 2008), 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; the avoidance of selection bias; and the impossibility of further data collection in historical contexts. Furthermore, as King (1999) notes, in some cases the bounds will be sufficiently informative, diagnostics will detect some violations of the modeling assumptions, and qualitative information may be included to improve the analysis. (The difficulties of ecological inference may become compounded in the EI-R framework, where ecological estimates are used as the dependent variable in second stage analyses in (Herron and Shotts, 2003b), and although extensions to the EI can alleviate some of these problems (Adolph and King, 2003; Herron and Shotts, 2003a; Adolph et al., 2003), there are still concerns over the use of the technique (Herron and Shotts, 2004; Cho and Gaines, 2004).)

Given the aforementioned difficulties of ecological inference and the fundamental lack of identification, an ecological analysis will benefit from a sensitivity analysis, and when possible, the inclusion of individual level data. In this paper, we discuss how to perform both of these tasks within the hierarchical convolution likelihood model described in Wakefield (2004a). The outline of this paper is as follows. In Section 2 we describe the fundamental difficulties of ecological inference, summarize some approaches to this inferential problem, and describe a data set that will be used to illustrate the various issues raised throughout the

paper. In Section 3 we describe the so-called convolution likelihood. Section 4 describes the Bayesian approach to inference, including subsections on prior specification and predictive distributions. The latter links the parameters of the model to the observables (unobserved counts), and is of great importance in political science applications (where there has been confusion between unobservable probabilities in a hypothetical super population model, and the samples fractions in the study population, which are potentially observable) as emphasized above. Section 5 describes the various computational schemes that have been suggested in the context of ecological inference. Section 6 presents a sensitivity analysis in the context of the Louisiana data, while we consider the combination of aggregate and individual data in Section 7. A concluding discussion rounds out the paper in Section 8.

2 The Fundamental Difficulty of Ecological Inference

To motivate our discussion we introduce a specific data set for which $Y = 0/1$ represents the event Democrat/Republican registration, and $X = 0/1$ the event Black/White. The data were collected in the U.S. state of Louisiana in 1990 and are available in each of the 64 counties of that state. These data are ideal for illustration of methods, since they are one of the few sources for which the individual level data are available.

Figure 1 gives an initial look at the ecological data. In panel (a) we give a histogram of the fraction black, and observe that in the majority of counties this fraction is less than 0.5. Panel (b) gives the proportion registered Republican against the fraction black, with a least squares line added to indicate the linear association. The general trend is that the proportion registered Republican decreases as the proportion black increases. The obvious explanation is that blacks are less likely to register Republican. Alternative explanations exist, however, in particular the same pattern could be observed if whites are less likely to register Republican if in a predominantly black county, if blacks are more likely to register Republican in a predominantly white county, or if individual race is an unimportant predictor of registration behavior, and instead an individual's behavior, whether black or white, is predicted by the proportion of blacks in the area. In each of these scenarios the proportion black/white in an area is an example of a *contextual* variable, a variable reflecting characteristics of individuals in a shared environment. To help explain the results that follow panels (c) and (d) give the fractions black and white who register Republican against the fraction black, again with least squares lines added. Note that it would not have been possible to produce these two plots without the individual level data which we have in this special case. We see that the fraction black who register Republican decreases across counties as the fraction black in the counties increases, hence it seems plausible that some contextual variable is driving black Republican registration (e.g. income). We also see that the proportion white who register Republican increases across counties as the fraction black in the counties increases. This is consistent with the explanation that whites in areas with large numbers of blacks are more fearful of affirmative action policies, and register/vote accordingly.

Hence Figure 1(b) understates the extent to which blacks are less likely to register Republican than whites, which is an example of what Selvin (1958), called the *ecological fallacy*: incorrect

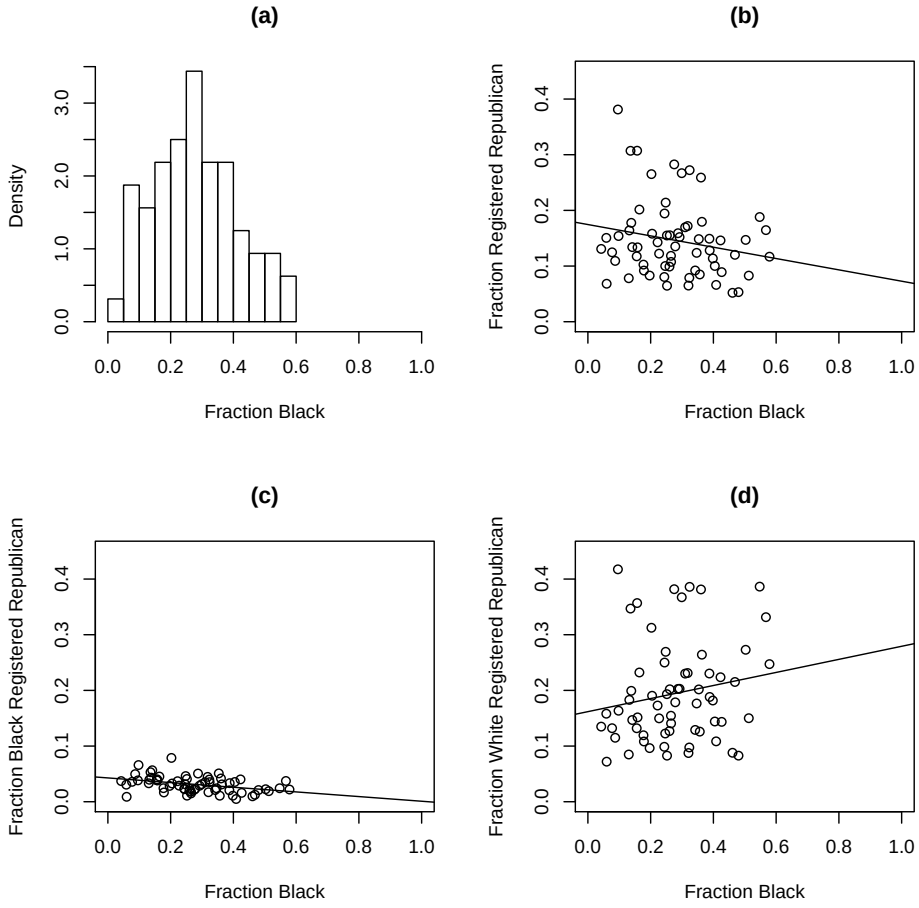

Figure 1: Across 64 counties of Louisiana: (a) Histogram of fraction of population that are black; (b) fraction registered Republican versus fraction black, (c) fraction black registered Republican versus fraction black, (d) fraction white registered Republican versus fraction black.

inference concerning individual effects gleaned from aggregate data. In an extreme case, the aggregate relationship could be the reversal of the true individual relationship, a phenomenon closely related to Simpson's paradox (Simpson, 1951), see Wakefield (2004c) for further discussion. The ecological fallacy had been discussed in the sociology literature before 1950, but Robinson (1950) provided an extremely lucid account, which explained the subsequent influence of the paper, in deterring the analysis of ecologic data. Recently, Robinson's paper has been revisited, within a multilevel framework (Subramanian et al., 2009b,a) and critiqued (Oakes, 2009; Firebaugh, 2009; Wakefield, 2009).

We now introduce some notation, in the context of the Louisiana data, in order to ease description of models that have been suggested for ecologic data. For a generic individual, $Y = 0/1$ will denote the event that an individual is unregistered/registered (the response), and $X = 0/1$ the event that an individual is of black/white race (the predictor). Table 1 describes the notation that we will use throughout the paper; Y_{0i}, Y_{1i} are the $Y = 1$ individuals from covariate group $X = 0, 1$, respectively, in area i. In an aggregate situation we do not observe the internal counts Y_{0i}, Y_{1i}. The fundamental difficulty of ecological inference is that we are interested in these two quantities, but it is their sum Y_i only, that we observe.

	$Y=0$	$Y=1$	
$x=0$		Y_{0i}	N_{0i}
$x=1$		Y_{1i}	N_{1i}
	$N_i - Y_i$	Y_i	N_i

Table 1: Table summarizing data in area i; in an ecological study the margins only are observed.

In the social science ecological inference literature the inference problem has often been treated as the imputation of the missing data, Y_{0i}, Y_{1i}, and due to this perspective, approaches have often implicitly adopted a finite sampling view. Here we utilize a hypothetical infinite population of exchangeable blacks and whites within each area as the primitive modeling object, and define the parameter p_{ji} to be the fraction of race j in area i that register. With this viewpoint an estimate of this probability, $\widehat{p}_{ji}$, is not equal to the true (but unobserved) fraction registered, Y_{ji}/N_{ji}, which we denote by $\widetilde{p}_{ji}$. In a finite sample view if Y_{ji} were observed then inference is complete since the population has been observed. In contrast, in the infinite population view, even if Y_{ji} is observed, uncertainty concerning p_{ji} will remain (though may be small if N_{ji} is large). However, while the infinite population model takes p_{ji} to be the primary parameter of interest, note that this model can still be used to make predictions about the fractions $\widetilde{p}_{ji}$ if these are of interest. Section 4.3 discusses this in greater detail.

To see the indeterminacy of ecological inference more clearly we write, for area i,

$$\frac{Y_i}{N_i} = \frac{Y_{0i} + Y_{1i}}{N_i} = \frac{Y_{0i}}{N_{0i}} \times \frac{N_{0i}}{N_i} + \frac{Y_{1i}}{N_{1i}} \times \frac{N_{1i}}{N_i}$$

which may be rewritten as

$$\widetilde{q}_i = \widetilde{p}_{0i} \times x_i + \widetilde{p}_{1i} \times (1 - x_i), \tag{1}$$

where $\widetilde{q}_i$ is the fraction registered, $\widetilde{p}_{0i}$ and $\widetilde{p}_{1i}$ are the black and white fractions registered, and x_i and $1 - x_i$ are the proportions black and white respectively. In an ecological data set, $\widetilde{q}_i$ and x_i are observed while $\widetilde{p}_{0i}$ and $\widetilde{p}_{1i}$ are not. From (1) we see that the observed $\widetilde{q}_i$ are consistent with many true fractions $\widetilde{p}_{0i}$, $\widetilde{p}_{1i}$. The bounds of Duncan and Davis (1953), may be written in terms of $\widetilde{q}_i$ and x_i:

$$\begin{aligned} \max\left\{0, \frac{\widetilde{q}_i - (1 - x_i)}{x_i}\right\} \leq \ \ \widetilde{p}_{0i} \ \ \leq \min\left\{1, \frac{\widetilde{q}_i}{x_i}\right\} \\ \max\left\{0, \frac{\widetilde{q}_i - x_i}{1 - x_i}\right\} \leq \ \ \widetilde{p}_{1i} \ \ \leq \min\left\{1, \frac{\widetilde{q}_i}{1 - x_i}\right\} \end{aligned}$$

In terms of the underlying probabilities p_{ji}, there is no constraint beyond $0 < p_{ji} < 1$. This is a crucial difference between the finite sample and infinite sampling views. Figure 2 shows the admissible ranges for $\widetilde{p}_{0i}$ and $\widetilde{p}_{1i}$ for the Louisiana data via a so-called *tomography plot*. We see that for the blacks in particular there is a great deal of uncertainty. The open circles correspond to the true fractions, which are available for these data.

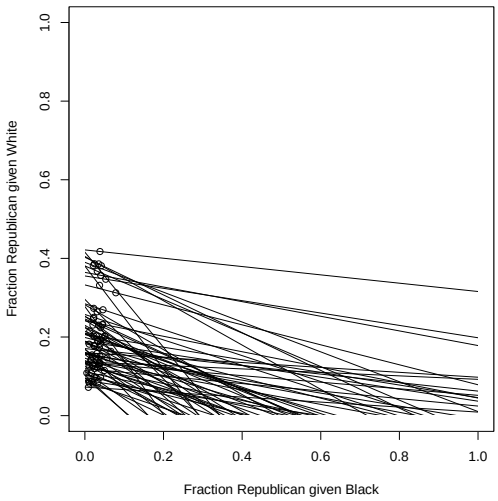

Figure 2: Tomography lines for Louisiana data.

Two extreme explanations are consistent with (1). First, following Goodman (1953, 1959) we may assume that $\widetilde{p}_{0i}$ and $\widetilde{p}_{1i}$ are such that

$$E[\widetilde{p}_{ji}|x_i] = p_j, \tag{2}$$

$j = 0, 1$ so that the fractions are uncorrelated with x_i. The expectation here is with respect to repeated sampling in areas with proportion of blacks x_i. We then have

$$E[\widetilde{q}_i|x_i] = p_0 \times x_i + p_1 \times (1 - x_i) = a + bx_i, \tag{3}$$

where $a = p_1$ and $b = p_0 - p_1$. Although it is only the expectations of the fractions that are considered constant in (2), the usual way of imputing the internal fractions is to simply take $\widetilde{p}_{ji} = p_j$, which is equivalent to a model in which the fractions themselves are constant. This model has sometimes been described as *Goodman regression*, but the name *ecological regression* is more appropriate as Goodman did not encourage general use of the approach, and in particular was aware that the 'constancy assumption' (2) would often be inappropriate. The assumption of constancy allows the mean to be derived, but to formulate an estimation method it would be desirable to derive the variance and covariance of $Y_i = N_i\widetilde{q}_i$. In general it has been assumed that counts in different areas are independent, and various forms for the variance have been considered. As we will describe in detail in Section 3, a plausible likelihood leads to Y_i following a convolution distribution with variance that depends on p_{0i} and p_{1i}.

A very simple model, termed the 'nonlinear neighborhood model' (Freedman et al., 1991), is to assume that $p_{0i} = p_{1i} = q_i$, i.e. to assume that registration and individual race are independent. This allows the table to be collapsed, and inference is straightforward. Freedman (2001) states that in this model, '...behavior is determined by geography not demography'. A specific version of the nonlinear neighborhood model, the 'linear neighborhood' model, was also described by Freedman et al. (1991) and makes the assumption that $E[p_{0i}|x_i]$ and $E[p_{1i}|x_i]$ are identical but depend on the proportion black via the linear form

$$E[p_{0i}|x_i] = E[p_{1i}|x_i] = E[q_i|x_i] = a + bx_i, \tag{4}$$

which is identical to (3) though the interpretation and imputed internal cells are drastically different under the two models, which was the motivation for Freedman et al. (1991) to introduce the model, to illustrate the fundamental unidentifiability of ecological inference. Other regression-type approaches, with a non-parametric flavor, are described by Chambers and Steel (2001).

The assumption that $\widetilde{q}_i$ is uncorrelated with x_i may be a major problem in some applications; see Freedman et al. (1998, 1999); Freedman (2001)) for examples. A further problem with ecological regression is the assumption that the *estimated* fractions are not allowed to vary across areas so that between-area variability is not acknowledged. Least squares procedures are known to provide consistent estimates of regression parameters under a range of distributions of the errors, but are also known to be very poor at providing *predictions* of observable quantities. For prediction some knowledge of the distribution of the error terms is required. The great benefit of the hierarchical approach that was popularized by King (1997) is that between-area differences in fractions are assigned a distribution, so allowing variability in the estimates of race-specific fractions across areas.

To conclude, in this section we have reviewed how two competing explanations with vastly different interpretations and inferential implications lead to an identical mean function. To overcome this unidentifiability and estimate $2m$ quantities from m observables, it is clear that any approach that is considered must make assumptions (or incorporate additional information). It is not immediately apparent, but also true that some of the assumptions from any approach will be uncheckable from the aggregate data alone. In all observational studies untestable assumptions such as 'no unmeasured confounding' are required for causal

interpretations (e.g. even Figures 1 (c) and (d) are not sufficient to derive the full causal story). If causal inference is the goal of an ecological study, this problem is particularly acute since the amount of information concerning quantities of interest is much smaller than in typical individual-level observational studies (e.g. Figure 1 (b) provides less information than Figures 1 (c) and (d)).

3 The Convolution Likelihood

In the previous section we simply derived the form of the marginal mean of the fraction registered Republican under various assumptions. In this section we describe a likelihood function under a plausible sampling scheme, and compare this with various (often implicit) likelihoods that have been used in the ecological literature. Recall that

$$p_{0i} = \Pr(Y = 1|x = 0, i) \quad \text{and} \quad p_{1i} = \Pr(Y = 1|x = 1, i) \tag{5}$$

are the population probabilities in area i, $i = 1, ..., m$. Returning to Table 1 we first note that if Y_{0i} and Y_{1i} were observed then if we were to assume that each of the N_{0i} black individuals in area i have independent Bernoulli responses with probability p_{0i}, and each of the N_{1i} white individuals in area i have independent Bernoulli response with probability p_{1i}, then

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

$j = 0, 1,\ i = 1, ..., m$. Under this sampling scheme, if Y_{0i} and Y_{1i} are unobserved then the sum Y_i follows a convolution of these binomial distributions:

$$\Pr(Y_i|p_{0i}, p_{1i}) = \sum_{y_{0i}=l_i}^{u_i} \binom{N_{0i}}{y_{0i}} \binom{N_{1i}}{Y_i - y_{0i}} p_{0i}^{y_{0i}}(1 - p_{0i})^{N_{0i}-y_{0i}} p_{1i}^{Y_i-y_{0i}}(1 - p_{1i})^{N_{1i}-Y_i+y_{0i}} \tag{6}$$

where

$$l_i = \max(0, Y_i - N_{1i}) \ , \qquad u_i = \min(N_{0i}, Y_i). \tag{7}$$

These values correspond to the admissible values that Y_i can take, given the margins in Table 1. McCullagh and Nelder (1989) consider this likelihood under the assumption that $p_{0i} = p_0$ and $p_{1i} = p_1$, see also Achen and Shively (1995, p. 46).

We now briefly examine the shape of the likelihood for a single table. 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. Steele et al. (2004) work with the convolution directly and report that the maximum likelihood estimator lies at the endpoint of the tomography line.

In King et al. (1999) the alternative model

$$Y_i|p_{0i}, p_{1i} \sim \text{Binomial}\{N_i, p_{0i}x_i + p_{1i}(1 - x_i)\} \tag{8}$$

was considered. As in King (1997), this produces a likelihood that is constant along the tomography line (an intuitively appealing feature given the implicit lack of information on

the internal cells of the table). However, the underlying individual-level model should be viewed as an approximation in this context since it assumes sampling *independently* N_i individuals each with probability $p_{0i}x_i + p_{1i}(1 - x_i)$.

In contrast, the convolution likelihood assumes that we sample N_{0i} individuals with probability p_{0i} and N_{1i} individuals with probability p_{1i}, $i = 1, ..., m$. As $N_i \to \infty$ with x_i and $\tilde{q}_i$ constant, this convolution likelihood function becomes concentrated along the tomography line with an asymmetric U-shape, with the maximum at one endpoint. At first, this non-constancy of the likelihood may seem counterintuitive (given the lack of information). However, an MLE on the boundary of the parameter space is an indication of a poorly behaved likelihood function. Furthermore, if uniform priors are placed on the probabilities p_{0i} and p_{1i}, this likelihood implies a flat posterior predictive distribution for $\widetilde{p}_{0i}$ and $\widetilde{p}_{1i}$ along the associated tomography line. Therefore, the convolution likelihood produces constancy for the predicted fractions ($\{\widetilde{p}_{0i}, \widetilde{p}_{1i}\}$ instead of $\{p_{0i}, p_{1i}\}$), but only produces constancy when the assumption of "no information" is made about p_{0i} and p_{1i}.

It is clear that the data in one table alone gives limited information concerning p_{0i}, p_{1i} or $\widetilde{p}_{0i}$, $\widetilde{p}_{1i}$, since we only have a single observation, Y_i. However, in most applications, ecological data from multiple areas is available.

4 Bayesian Inference

4.1 Priors

Following King (1997) a number of authors have developed hierarchical approaches in which, rather than reduce the dimensionality of the models as was described in the previous section, the full $2m$ parameters are retained but the probabilities/fractions, are assumed to arise from a bivariate distribution.

At the second stage of the King (1997) model it is assumed that the pair $\tilde{p}_{0i}, \tilde{p}_{1i}$ arise from a truncated bivariate normal distribution, hence imposing identifiability. King (1997) views the truncated bivariate normal distribution as the likelihood while we have referred to the tomography lines as providing the first stage of the model, with the truncated bivariate normal the second stage of the model. Inference is initially carried out via MLE for the five population parameters, using numerical integration, and then simulation is used to make more refined inference. Priors may be placed on the population parameters (that characterize the truncated normal) to give a Bayesian model. In common with the majority of approaches, it is assumed that the pairs of fractions form an independent sample from the second stage distribution (here the truncated bivariate normal), see Haneuse and Wakefield (2004) for a hierarchical model with spatial dependence between the probabilities. The model in its most basic form also assumes that the fractions are uncorrelated with x_i. The latter may be relaxed (see King 1997, Chapter 9), via the introduction of contextual effects (in King et al. (2004) it is recommended that such effects be included), but reliable estimation of both individual and contextual effects is crucially dependent on the existence of substantive

prior information (see the example in Wakefield (2004c), for a further demonstration of this). The freely-available EzI software (Benoit and King, 1998) may be used to implement the truncated normal model, and its extensions.

At the second stage, King et al. (1999) assume that p_{0i} and p_{1i} are independent with

$$p_{ji}|a_j, b_j \sim_{iid} \text{Beta}(a_j, b_j). \tag{9}$$

The third and final stage of the model consists of exponential priors, Exp(λ) on a_j, b_j, $j = 0, 1$, where λ^{-1} is the mean of the exponential. Specifically, in the example considered it was assumed that these exponential priors had mean 2 ($\lambda = 0.5$), a choice which may not be desirable in many instances because it often produces a prior for each probability which is very strongly U-shaped (since beta priors with $a_j < 1, b_j < 1$ are themselves U-shaped, and an exponential with mean 2 has a 0.39 probability of being less than one). This is discussed more fully in Wakefield (2004c), in particular see Figure 6. Choosing much smaller values of λ, for example, $\lambda = 0.01$, produces almost uniform priors on the probabilities, though we would not universally recommend a particular hyperprior, given the sensitivity of inference it should be context specific. As the number of tables decreases and the x distribution becomes more asymmetric this problem becomes more and more acute. The ideal situation is for substantive information to be available for prior specification. The strong dependence on the third stage prior is in stark contrast to the usual generalized mixed model case for which there is far less dependence (except for priors on variance components, where again care must be taken with small numbers of units). Here we emphasize that the form of the prior should be examined through simulation. Specifically, for generic second stage, $p(p|\phi$, and third stage, $p(\phi)$:

1. For fixed ϕ, simulate $\phi^{(s)} \sim p(\phi)$, for $s = 1, ..., S$.
2. Simulate $p^{(s)} \sim p(p|\phi^{(s)})$, for $s = 1, ..., S$.
3. Examine graphical and numerical summaries of the collection $\{\phi^{(s)}, s = 1, ..., S\}$.

This procedure will be illustrated shortly.

The model given by (9) does not allow dependence between the two random effects (note this is distinct from the independence between pairs of random effects in different areas, which is also assumed) though it is conjugate (giving a marginal distribution for the data that is beta-binomial) which may offer some advantage in terms of computation. The model also allows area-level covariates to be added at the second stage.

Wakefield (2004c) proposed, as an alternative to the beta model, a second stage in which the logits of the registration probabilities arose from a bivariate normal distribution; this model was introduced for the analysis of a series of 2×2 tables when the internal cells were observed by Skene and Wakefield (1990). Specifically, a reasonably general form is

$$\begin{aligned} \theta_{0i} &= \log\left(\frac{p_{0i}}{1-p_{0i}}\right) = \mu_0 + \beta_0 z_i + \delta_{0i} \\ \theta_{1i} &= \log\left(\frac{p_{0i}}{1-p_{0i}}\right) = \mu_1 + \beta_1 z_i + \delta_{1i} \end{aligned} \tag{10}$$

with

$$\delta_i \sim N_2(0, \Sigma),$$

where

$$\delta_i = \begin{bmatrix} \delta_{0i} \\ \delta_{1i} \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{bmatrix}. \tag{11}$$

Hence θ_{0i} and θ_{1i} denote the logits of the probabilities p_{0i} and p_{1i} in table i, so that $p_{ji} = \exp(\theta_{ji})/\{1 + \exp(\theta_{ji})\}$, $j = 0, 1$. In the specification (10), z_i represent area-level characteristics (and may, in principle, include x_i) and β_0, β_1 are (ecological) log odds ratios associated with these variables.

A third stage hyperprior adds priors on μ_0, μ_1 and Σ (and β_0, β_1 if there are covariates). It is difficult to gain information on the covariance term Σ_{01} and so from this point onwards we assume that $\Sigma_{01} = 0$. Without substantive information for the registration-race data, Wakefield (2004c) chose logistic priors with location 0 and scale 1 for μ_0 and μ_1, since these induce uniform priors on $\exp(\mu_j)/\{1 + \exp(\mu_j)\}$ (the median of the registration probability for race j across the population of areas). Since $G(z) \approx (cz)$ with $c = 16\sqrt{3}/(15\pi)$ and where $G(z) = (1 + e^{-z})^{-1}$ is the CDF of a logistic random variables, as an alternative we may specify normal priors with mean 0 and standard deviation $1/c$.

For the precisions $\Sigma_{00}^{-1}, \Sigma_{11}^{-1}$ we specify gamma distributions $\text{Ga}(a, b)$ (where the parameterization is such that the mean is given by a/b). In the `WinBUGS` manual the priors Ga(0.001, 0.001) are often used for precisions within a hierarchical model. This choice is not to be recommended in general (that is, for all applications); here it is a very poor one (and leads to marginal priors for the probabilities that are highly U-shaped). We follow a previously suggested procedure (Wakefield, 2009) which we briefly describe for a generic log odds ratio in area i, $\delta_i \sim_{iid} N(0, \Sigma)$ with $\Sigma^{-1} \sim \text{Ga}(a, b)$. We integrate over Σ to find the marginal distribution $p(\delta_i)$ which is a t distribution with $d = 2a$ degrees of freedom, location zero, and scale $\Sigma = b/a$. To construct a prior distribution we require a careful interpretation of δ_i, or more informatively, $\exp(\delta_i)$ which is the perturbation of the odds of Republican registration from the median of the distribution of the odds of Republican registration across all areas. Hence we may refer to $\exp(\delta_i)$ as a residual odds, since it is relative to the median odds across areas. We give a range for $\exp(\delta_i)$. In particular, for the range $(1/R, R)$ we use the relationship $\pm t^d_{0.025}\sqrt{\sigma} = \pm \log R$, where t^d_r is the $100 \times r$-th quantile of a Student t random variable with d degrees of freedom, to give $a = d/2$, $b = (\log R)^2 d/2(t^d_{1-(1-q)/2})^2$. We choose $d = 1$, to give a Cauchy marginal distribution. As an example, for a 95% range of $[0.1, 10]$ we obtain $a = 0.5, b = 0.0164$.

We illustrate the prior simulation strategy with the priors $\mu \sim N(0, 1/c^2)$, $\Sigma^{-1} \sim \text{Ga}(0.5, b)$, with $\theta_i = \mu + \delta_i$, and $\delta_i \sim N(0, \Sigma)$. We take three values of b, corrsponding to 95% ranges of $[0.9, 1.1]$, $[0.5, 2]$ and $[0.1, 10]$ (which correspond to $b = 0.000028, 0.00149, 0.0164$, respectively). Figure 3(a) gives the marginal distribution of $\text{median}(p) = \frac{\exp(\mu)}{1+\exp(\mu)}$, which is close to uniform, in line with the theory outlined above. These priors are applied in Section 6.

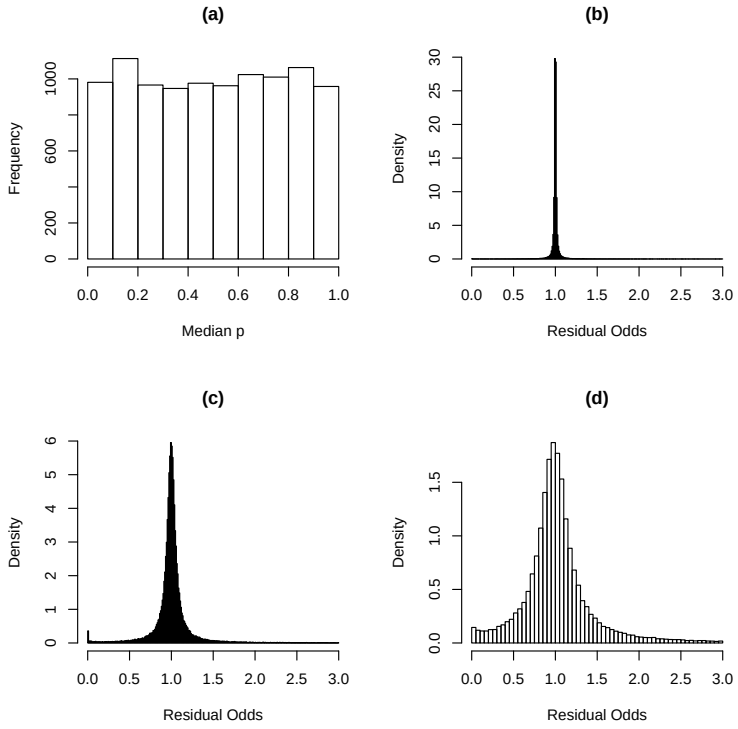

Figure 3: Simulations from the $N(0, 1/c^2) \times \text{Ga}(0.5, b)$ prior. Panel (a) gives the marginal distribution of the median odds of Republican registration. Panels (b), (c) and (d) gives the residual odds of Republican registration (across areas) with ranges $[0.9, 1.1]$, $[0.5, 2]$ and $[0.1, 10]$, respectively.

4.2 Derivation of the posterior distribution

In the Bayesian approach all unknown quantities are assigned prior distributions and the posterior distribution reflects both these distributions and the information in the data that is contained in the likelihood. In the hierarchical models described in Section 4.1 two stage priors are specified, with the first stage of the prior assuming a common form for the pairs of probabilities, and the second stage assigning *hyperpriors* to the parameters of this form. Letting $\boldsymbol{p}_i$ represent the pair of table specific probabilities, and $\boldsymbol{\phi}$ a generic set of hyperparameters upon which the second stage of the prior depends, we have:

$$\pi(\boldsymbol{p}_1, ..., \boldsymbol{p}_m, \boldsymbol{\phi}|y_1, ..., y_m) \propto p(y_1, ..., y_m|\boldsymbol{p}_1, ..., \boldsymbol{p}_m, \boldsymbol{\phi}) \times \pi(\boldsymbol{p}_1, ..., \boldsymbol{p}_m, \boldsymbol{\phi})$$

with

$$p(y_1, ..., y_n|\boldsymbol{p}_1, ..., \boldsymbol{p}_m, \boldsymbol{\phi}) = \prod_{i=1}^{m} p(y_i|\boldsymbol{p}_i),$$

by conditional independence of counts in different areas, and

$$\pi(\boldsymbol{p}_1, ..., \boldsymbol{p}_m, \boldsymbol{\phi}) = \pi(\boldsymbol{p}_1, ..., \boldsymbol{p}_m|\boldsymbol{\phi}) \times \pi(\boldsymbol{\phi}),$$

to give the two-stage prior. Under the assumption of independence of the table-specific parameters (which would not be true if we assumed spatial dependence between these parameters), we may further write

$$\pi(\boldsymbol{p}_1, ..., \boldsymbol{p}_m|\boldsymbol{\phi}) = \prod_{i=1}^{m} \pi(\boldsymbol{p}_i|\boldsymbol{\phi}).$$

Hence, under these assumptions, we have the posterior distribution

$$\pi(\boldsymbol{p}_1, ..., \boldsymbol{p}_m, \boldsymbol{\phi}|y_1, ..., y_m) \propto \prod_{i=1}^{m} p(y_i|\boldsymbol{p}_i) \times \prod_{i=1}^{m} \pi(\boldsymbol{p}_i|\boldsymbol{\phi}) \times \pi(\boldsymbol{\phi}).$$

Inference follows via consideration of marginal posterior distributions, and predictive distributions. For example $\pi(\boldsymbol{p}_i|y_1, ..., y_m)$ is the marginal posterior distribution for the pair of probabilities of Republican registration from table i.

4.3 The Posterior Predictive Distribution

We may also be interested in imputing the missing counts in area i. In particular, this is often the goal of ecological inference in the social sciences. This type of inference may be carried out via examination of the predictive distribution

$$\Pr(Y_{0i}|y_1, ..., y_m) = \int \Pr(Y_{0i}|\boldsymbol{p}_i, N_{0i}, N_{1i}, N_i - y_i, y_i) \times \pi(\boldsymbol{p}_i|y_1, ..., y_n) d\theta_i.$$

Note that we only need the distribution for Y_{0i} since the distribution of $Y_{1i} = Y_i - Y_{0i}$, which is immediately available. The integral averages the distribution of $\Pr(Y_{0i}|\boldsymbol{p}_i, N_{0i}, N_{1i}, N_i - y_i, y_i)$ with respect to the posterior. The distribution of Y_{0i} given the row and column

margins and the table probabilities, is a non-central (sometimes referred to as an extended) hypergeometric distribution, see for example, McCullagh and Nelder (1989). Suppose the odds ratio in the table is given by $\psi_i = p_{0i}(1-p_{1i})/p_{1i}(1-p_{0i})$; then Y_{0i} has a non-central hypergeometric distribution if its distribution is of the form

$$Pr(Y_{0i} = y_{0i}|\psi_i, N_{0i}, N_{1i}, N_i - y_i, y_i) = \begin{cases} \dfrac{\binom{N_{0i}}{y_{0i}}\binom{N_{1i}}{y_i - y_{0i}}\psi_i^{y_{0i}}}{\sum_{u=l_i}^{u_i}\binom{N_{0i}}{u}\binom{N_{1i}}{y_i - u}\psi_i^{u}} & y_{0i} = l_i, ..., u_i, \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

where $l_i = \max(0, y_i - N_{1i})$ and $u_i = \min(N_{0i}, y_i)$. Hence the predictive distribution is an overdispersed non-central hypergeometric distribution. The above predictive distribution produces $(y_{0i}/N_{0i}, y_{1i}/N_{1i})$ pairs that lie along the tomography line, and with flat priors on the probabilities, this distribution is uniform along the tomography line Wakefield (2004d). This provides a link with the likelihoods of King (1997) and King et al. (1999), but we emphasize that the flat distribution is with respect to the fractions, and not for the underlying probabilities.

5 Computation

Given the lack of identifiability in the posterior distribution, it is not surprising that computation is not straightforward for ecological inference, when analyzed using hierarchical models. In Wakefield (2004c) an "obvious" augmented data scheme was utilized.

Auxiliary Variable Sampling:

For the missing data y_{0i}, the distribution is an extended hypergeometric distribution with margins $N_{0i}, N_{1i}, y_i, N_i - y_i$, $i = 1, ..., m$:

$$\Pr(Y_{0i} = y_{0i}|\psi_i, N_{0i}, N_{1i}, y_i) = \begin{cases} \dfrac{\binom{N_{0i}}{y_{0i}}\binom{N_{1i}}{y_i - y_{0i}}\psi_i^{y_{0i}}}{\sum_{u=l_i}^{u_i}\binom{N_{0i}}{u}\binom{N_{1i}}{y_i - u}\psi_i^{u}} & y_{0i} = l_i, ..., u_i, \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

where $l_i = \max(0, Y - N_{1i})$ and $u_i = \min(N_{0i}, Y)$ and $\psi_i = p_{0i}(1-p_{1i})/p_{1i}(1-p_{0i})$ 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. The mode is available in closed form, however, which may be exploited to produce an improved scheme, see Wakefield (2004c) for details.

Posterior Probability Sampling:

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

$$\text{Be}(y_{0i} + a_0, N_{0i} - y_{0i} + b_0)\text{Be}(y_{1i} + a_1, N_{1i} - y_{1i} + b_1), \tag{14}$$

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 large table Wakefield (2004c) proposed a normal approximation to the convolution. `WinBUGS` code for ecological inference using this normal approximation was given in the Appendix of Wakefield (2004b). In the JAGS (Just Another Gibbs Sampler) software (Plummer (2009)) there is a a novel distribution `dsum` that may be used in the ecological inference context. It may be used in the following way (Plummer, personal communication). The specification `y` $\sim$ `dsum(y0, y1)` where `y` is observed and `y0, y1` are unobserved discrete-valued stochastic nodes creates an MCMC sampler that will simultaneously update `y0` and `y1`, while respecting the constraint `y0 + y1== y`. In typical social science applications there are too many possible values of `y0 (or y1)` to use inversion and so the sampler uses discrete slice sampling (Neal, 2003) as an alternative.

The `R` package `MCMCpack` package (Martin et al., 2009) contains a function `MCMChierEI` to implement the hierarchical model of Wakefield (2004c) with the normal approximation to the convolution likelihood implemented along with slice sampling. An extension to this work to the $R \times C$ table case is available in the `R` package `RxCEcolInf` (Greiner et al., 2009). This package contains functions to analyze both ecological data alone, or ecological data supplemented with individual-level data, which is a very important extension, as we discuss further in Section 7. The methodological extension to the 2×2 table cases is described in Greiner and Quinn (2009). An additional `R` package, `eco` is also available and fits models described in Imai et al. (2008). These models include a close relative of the parametric model of Wakefield (2004c) and a non-parametric model in which the second stage distribution is a Dirichlet process prior.

With sampling-based inference, if we can simulate from $\Pr(Y_{0i}, Y_{1i}|\theta_i, N_{0i}, N_{1i}, N_i - y_i, y_i)$ then it is straightforward to simulate from the predictive distribution, once samples $\boldsymbol{p}_i^{(s)}$ are available from $\pi(\boldsymbol{p}_i|N_{0i}, N_{1i}, N_i - y_i, y_i)$, via

$$\frac{1}{S}\sum_{s=1}^{S}\Pr(Y_{0i}, Y_{1i}|\boldsymbol{p}_i^{(s)}, N_{0i}, N_{1i}, N_i - y_i, y_i).$$

6 Illustrative Analysis

We analyze the Louisiana data using the hierarchical normal model, and investigate the sensitivity of inference to the choice of the gamma prior on the precisions of the random effects distribution. Specifically, following the procedure outlined in Section 4.1 we pick ranges of: $[0.1, 10]$, $[0.2, 5]$, $[0.5, 2]$, $[0.9, 1.1]$, for the residual odds of Republican registration. These range from a prior that expexts the probabilities across areas to be tightly clustered

around the median, to one in which there is much larger variability. In all cases we specify $N(0, (15\pi/16\sqrt{3})^2)$ priors for μ_j, $j = 1, 2$.

We used the `MCMChierEI` function to carry out inference, and ran the Markov chains for 10^6 iterations, after discarding 10^5 iterations as burn-in. We summarize the accuracy of inference in terms of S_0 and S_1 where $S_j = \sum_i |\hat{p}_{ji} - \tilde{p}_{ji}|/\tilde{p}_{ji}$, $j = 0, 1$. For the black probabilities we obtain $S_0 = 76.1, 72.8, 78.9, 81.6$ while for the white probabilities $S_1 = 4.4, 4.2, 4.6, 4.7$. The first thing to note is that inference for the black proportions is much less accurate, as we might expect from Figure 2, since there is far less information available. The empirical distribution of the residual odds may be calculated here (since the individual data are available), and give 95% ranges for blacks and whites of [0.33,2.3] and [0.41,2.9], respectively so that the second prior is most consistent with the data, which explains the above summaries of S_0 and S_1. Figure 4 gives the posterior medians of the fractions registered Republican for blacks and

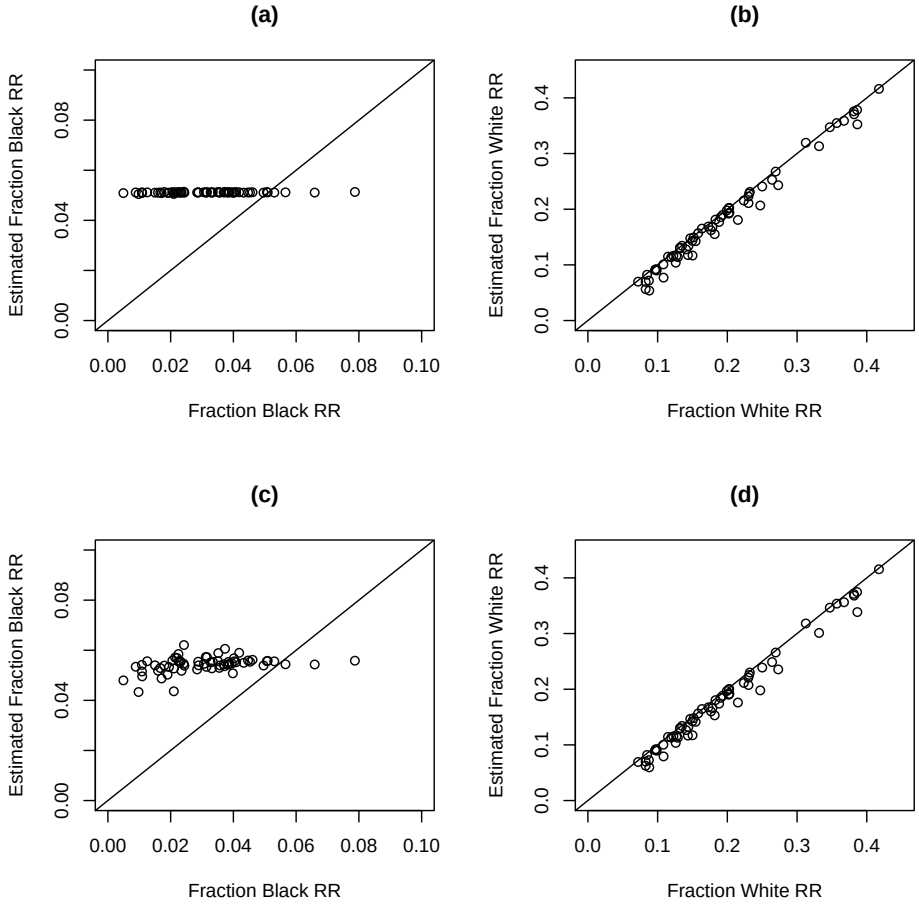

Figure 4: (a) Estimated black fraction registered Republican (RR) versus black fraction RR under the narrow prior, (b) estimated white fraction RR versus white fraction RR under the narrow prior, (c) estimated black fraction RR versus black fraction RR under the wide prior, (d) estimated white fraction RR versus white fraction RR under the wide prior.

Table 2: Summary of notation for the situation in which we have both individual survey data with sample sizes $m_0 i$ and m_{1i}, and aggregate marginal data in area i There are N_i individuals in area i, with y_i responding $Y = 1$, and N_{0i}, N_{1i} individuals with $x = 0, 1$ respectively.

	Survey Data			Aggregate Data		
	$Y=0$	$Y=1$		$Y=0$	$Y=1$	
$x=0$		z_{0i}	m_{0i}			$N_{0i}-m_{0i}$
$x=1$		z_{1i}	m_{1i}			$N_{1i}-m_{1i}$
	$m_i - z_i$	z_i	m_i	$N_i - m_i - (y_i - z_i)$	$y_i - z_i$	$N_i - m_i$

whites, versus the true fractions based on the individual data. The top row is under the prior with residual odds in the range [0.9,1.1] and the bottom row is under the prior with range [0.1,10]. We see that the black/white fractions tend to be overestimated/underestimated. The effect of the prior is most apparent for the black fractions; under the narrower prior the estimates are virtually identical for all counties. In Figure 2 we saw that in many counties the bounds on the black fractions were wide, indicating the lack of information.

7 Combination of Individual and Aggregate Data for the Posterior Predictive

We now consider the situation in which survey data are available, Table 2 illustrates the notation in this case, the observed counts in area i 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 combined with the aggregate data likelihood, with each term being independent, i.e.

$$L(p_{0i}, p_{1i}) = p(z_{0i}|p_{0i}) \times p(z_{1i}|p_{1i}) \times p(y_i^\star|p_{0i}, p_{1i}),$$

where $y_i^\star = y_i - z_i$, and each of the first two terms is binomial and the third is the convolution likelihood.

Wakefield (2004c) illustrated the benefits of adding individual (survey) data to the ecological data to gain identifiability. A number of discussants to the paper (Best, 2004; Jackson, 2004; Salway, 2004) and a subsequent paper (Glynn et al., 2008) suggested that smaller sample sizes may be all that is needed, and that the design of the survey is an important topic. Here we touch upon these issues by investigating a number of scenarios.

The first 4 rows of Table 3 report S_0 and S_1 for survey samples within each area of sizes 1000, 500, 300, 100, respectively. Two sets of results are given in each row, the first set correspond to the use of the individual-level data only, and the second to the combined individual-ecologic data. All results were obtained using the `AnalyzeWithExitPoll` function within the `RxCEcolInf` package. Each MCMC run began with a burn-in of 10^5, with 10^3 samples collected subsequently over 10^6 iterations. In all cases inference is greatly improved when

Data Source	Individual S_0	Individual S_1	Combined S_0	Combined S_1
Ecologic Only	—	—	78.9	4.6
1000 Samples	22.1	4.0	16.7	0.9
500 Samples	26.2	6.1	15.9	0.8
300 Samples	40.4	8.1	21.7	1.1
100 Samples	73.1	13.4	31.1	1.7
Samples in Half	—	—	22.0	1.3
Samples in Quarter	—	—	24.9	1.6
Samples in Eighth	—	—	29.5	1.6

Table 3: Summaries for the combined survey/ecologic setting. Rows two through five of the main body show the effects of adding samples of the stated sizes to all areas. Error measures associated with the individual data only are also given for these rows. The last three rows report situations in which individual samples of size 500 were sampled from the reported proportion of areas.

individual-level data supplement the ecologic data. Examination of the resultant estimates of the probabilities revealed that for low sample sizes bias existed in the estimates but this was more than offset by the reduction in variance.

Viewed in the opposite direction to the emphasis of this paper, an important observation is that inference based on the individual data only can be greatly improved by supplementing the survey data with ecologic information.

In the second stage of our investigation we now fixed the sample size at 500 but only sampled 1/2, 1/4 and 1/8 of areas. Again we see that in all cases inference is hugely improved over the ecologic only analysis. In fact, note that with samples of size 500 from 8/64 tables (4000 sample size), we achieve better results than with samples of 300 from 64/64 tables (19,200 sample size). Given the additional cost effectiveness of sampling within fewer tables, this result implies the potential for considerable savings from sampling design in this context.

In Figure 5 we plot the estimates versus the "true" fractions. In the top row these comparisons are for the areas with survey data, while in the second row the comparisons are in areas with no survey data. Although there is clearly bias in the estimates for blacks in particular, it is far less than in the ecologic only case (compare with Figure 4). This improvement in estimation with no survey data is due to the hierarchical model which is common to all areas, thus allowing the areas with surveys to positively impact the areas with no data.

8 Discussion

In this paper, we have shown that it can be unreliable to estimate the p_{ji} or $\widetilde{p}_{ji}$ values with ecological data. We have also shown that the inclusion of individual level data in the analysis can mitigate these problems, and that only a small amount of data from a small number of

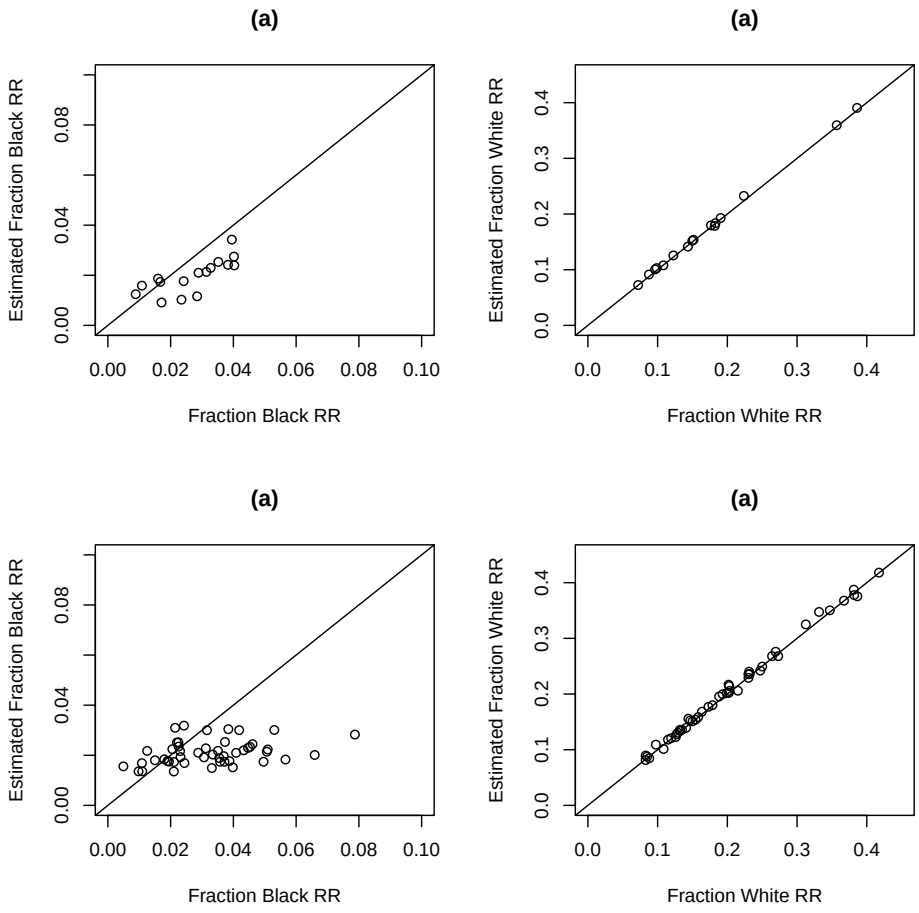

Figure 5: Analysis based on subsamples in the first 16 areas only: (a) Estimated black fraction registered Republican (RR) versus black fraction RR in areas with survey data, (b) estimated white fraction RR versus white fraction RR in areas with survey data, (c) estimated black fraction RR versus black fraction RR in areas with no survey data, (d) estimated white fraction RR versus white fraction RR in areas with no survey data.

tables may be necessary.

Furthermore, while we have focused on the estimation of p_{ji} or $\widetilde{p}_{ji}$ in this paper, it is straightforward to see that the estimation of contextual parameters is not possible from ecological data alone. Let p_{ji} be the proportion of individuals of race j in area i and suppose the individual-level model is:

$$p_{ji} = a_{ji} + b_{ji}x_i$$

so that we have both effects due to race in area i, a_{0i} and a_{1i}, and contextual effects, b_{0i} and b_{1i}. Upon averaging across individuals, to give ecological data we obtain the marginal area-level probability

$$p_i = x_i \times p_{0i} + (1 - x_i)p_{1i} = a_{1i} + x_i(a_{0i} + b_{1i} - a_{1i}) + x_i^2(b_{0i} - b_{1i})$$

which clearly shows the identifiability problem. With a non-linear model (for example a logistic model), the parameters become identifiable, but only due to the non-linearity and different non-linear forms will give different answers.

Data and code for all of the examples of the paper are available at:

`http://faculty.washington.edu/jonno/cv.html`

In Section 7 we showed the benefits of small amounts of individual-level data. In the simulations we sampled individuals without replacement from black and white populations to produce a representative sample. In practice selection bias may be present, and without further information on the nature of this bias, a sensitivity analysis should be performed that incorporates the uncertainty about selection effects. However, within the framework we have described such a study is straightforward.

Acknowledgements

The authors would like to thank Kevin Quinn, Sebastien Haneuse, and Gary King for useful discussions. The second author was supported by grant R01 CA095994 from the National Institutes of Health.

References

Achen, C. H. and W. P. Shively (1995). *Cross-level Inference*. University of Chicago Press.

Adolph, C. and G. King (2003). Analyzing Second-Stage Ecological Regressions: Comment on Herron and Shotts. *Political Analysis 11*(1), 65–76.

Adolph, C., G. King, M. Herron, and K. Shotts (2003). A Consensus on Second-Stage Analyses in Ecological Inference Models. *Political Analysis 11*(1), 86–94.

Benoit, K. and G. King (1998). *EzI: An easy program for ecological inference, version 2.02*. Boston: Department of Government, Harvard University.

Best, N. (2004). Discussion of: "ecological inference for 2 × 2 tables". *Journal of the Royal Statistical Society, Series A 167*, 426–427.

Burden, B. and D. C. Kimball (1998). A New Approach to the Study of Ticket Splitting. *American Political Science Review 92*(3), 533–544.

Chambers, R. and D. G. Steel (2001). Simple methods for ecological inference in 2×2 tables. *Journal of the Royal Statistical Society Series A 164*, 175–92.

Cho, W. K. T. (1998). Iff the Assumption Fits...: A Comment on the King Ecological Inference Solution. *Political Analysis 7*(1), 143–163.

Cho, W. K. T. and B. J. Gaines (2004). The Limits of Ecological Inference: The Case of Split-Ticket Voting. *American Journal of Political Science 48*(1), 152–171.

Cleave, N., P. J. Brown, and C. D. Payne (1995). Evaluation of methods for ecological inference. *Journal of the Royal Statistical Society, Series A 158*, 55–72.

Duncan, O. D. and B. Davis (1953). An alternative to ecological correlation. *American Sociological Review 18*, 665–6.

Firebaugh, G. (2009). Commentary: 'is the social world flat? W.S. Robinson and the ecologic fallacy'. *International Journal of Epidemiology 38*, 368–370.

Freedman, D. (2001). Ecological inference and the ecological fallacy. In N. Smelser and P. Baltes (Eds.), *International Encyclopedia of the Social and Behavioural Sciences, Volume 6*, pp. 4027–4030. New York: Elsevier.

Freedman, D. A., S. P. Klein, M. Ostland, and M. R. Roberts (1998). A solution to the ecological inference problem (book review). *Journal of the American Statistical Association 93*, 1518–1522.

Freedman, D. A., S. P. Klein, J. Sacks, C. A. Smyth, and C. G. Everett (1991). Ecological regression and voting rights. *Evaluation Review 15*, 673–711.

Freedman, D. A., M. Ostland, M. R. Roberts, and S. P. Klein (1999). Reply to g. king. *Journal of the American Statistical Association 94*, 355–357.

Gelman, A., D. Park, S. Ansolabehere, P. Price, and L. Minnite (2001). Models, assumptions and model checking in ecological regressions. *Journal of the Royal Statistical Society, Series A 164*, 101–118.

Glynn, A., J. Wakefield, M. Handcock, and T. Richardson (2008). Alleviating linear ecological bias and optimal design with subsample data. *Journal of the Royal Statistical Society, Series A 171*, 179–202.

Goodman, L. A. (1953). Ecological regressions and the behavior of individuals. *American Sociological Review 18*, 663–4.

Goodman, L. A. (1959). Some alternatives to ecological correlation. *Americal Journal of Sociology 64*, 610–25.

Greenland, S. and J. Robins (1994). Ecological studies: biases, misconceptions and counterexamples. *American Journal of Epidemiology 139*, 747–760.

Greiner, D., P. Baimnes, and K. Quinn (2009). *Package 'RxCEcolInf'*.

Greiner, D. and K. Quinn (2009). $r \times c$ ecological inference: bounds, correlations, flexibility and transparency of assumptions. *Journal of the Royal Statistical Society, Series A 172*, 67–81.

Haneuse, S. and J. Wakefield (2004). Ecological inference incorporating spatial dependence. In G. King, O. Rosen, and M. Tanner (Eds.), *Ecological Inference: New Methodological Strategies*, Chapter 12, pp. 266–302. Cambridge: Cambridge University Press.

Herron, M. and K. Shotts (2003a). Cross-Contamination in EI-R: Reply. *Political Analysis 11*(1), 77–85.

Herron, M. and K. Shotts (2003b). Using Ecological Inference Point Estimates as Dependent Variables in Second-Stage Linear Regressions. *Political Analysis 11*(1), 44–64.

Herron, M. and K. Shotts (2004). Logical Inconsistency in EI-Based Second-Stage Regressions. *American Journal of Political Science 48*(1), 172–183.

Imai, K., Y. Lu, and A. Strauss (2008). Bayesian and likelihood inference for 2×2 ecological tables: an incomplete data approach. *Political Analysis 16*, 41–69.

Jackson, C. (2004). Discussion of: "ecological inference for 2 × 2 tables". *Journal of the Royal Statistical Society, Series A 167*, 430.

King, G. (1997). *A Solution to the Ecological Inference Problem.* Princeton: Princeton University Press.

King, G. (1999). The Future of Ecological Inference Research: A Comment on Freedman et al. *Journal of the American Statistical Association 94*(445), 352–355.

King, G., O. Rosen, and M. Tanner (2004). Information in ecological inference: An introduction. In G. King, O. Rosen, and M. Tanner (Eds.), *Ecological Inference: New Methodological Strategies*, pp. 1–12. Cambridge: Cambridge University Press.

King, G., O. Rosen, and M. A. Tanner (1999). Binomial-beta hierarchical models for ecological inference. *Sociological Methods and Research 28*, 61–90.

Lasserre, V., C. Guihenneuc-Jouyaux, and S. Richardson (2000). Biases in ecological studies: utility of including within-area distribution of confounders. *Statistics in Medicine 19*, 45–59.

Martin, A., K. Quinn, and J. Park (2009). *Package 'MCMCpack'*.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models, Second Edition*. London: Chapman and Hall.

Neal, R. (2003). Slice sampling (with discussion). *Annals of Statistics 31*, 705–767.

Oakes, J. (2009). Commentary: Individual, ecological and multilevel fallacies. *International Journal of Epidemiology 38*, 361–368.

Plackett, R. (1977). The marginal totals of a 2×2 table. *Biometrika 64*, 37–42.

Plummer, M. (2009). Jags version 1.0.3 manual. Technical report.

Plummer, M. and D. Clayton (1996). Estimation of population exposure. *Journal of the Royal Statistical Society, Series B 58*, 113–126.

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

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

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

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review 15*, 351–57.

Salway, R. (2004). Discussion of: "ecological inference for 2×2 tables". *Journal of the Royal Statistical Society, Series A 167*, 438–439.

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

Selvin, H. (1958). Durkheim's 'suicide' and problems of empirical research. *American Journal of Sociology 63*, 607–619.

Simpson, E. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B 13*, 238–241.

Skene, A. and J. Wakefield (1990). Hierarchical models for multi-centre binary response studies. *Statistics in Medicine 9*, 919–929.

Steele, D., E. Beh, and R. Chambers (2004). The information in aggregate data. In G. King, O. Rosen, and M. Tanner (Eds.), *Ecological Inference: New Methodological Strategies*. Cambridge: Cambridge University Press.

Subramanian, S., K. Jones, A. Kaddour, and N. Krieger (2009a). Response: The value of a historically informed multilevel analysis of Robinson's data. *International Journal of Epidemiology 38*, 370–373.

Subramanian, S., K. Jones, A. Kaddour, and N. Krieger (2009b). Revisiting Robinson: the perils of individualistic and ecologic fallacy. *International Journal of Epidemiology 38*, 342–360.

Wakefield, J. (2004a). Ecological inference for 2 x 2 tables (with discussion). *Journal of the Royal Statistical Society, Series A 167*, 385–445.

Wakefield, J. (2004b). Prior and likelihood choices in the analysis of ecological data. In G. King, O. Rosen, and M. Tanner (Eds.), *Ecological Inference: New Methodological Strategies*, Chapter 1, pp. 13–50. Cambridge: Cambridge University Press.

Wakefield, J. (2008). Ecologic studies revisited. *Annual Review of Public Health 29*, 75–90.

Wakefield, J. (2009). Multi-level modelling, the ecologic fallacy, and hybrid study designs. *International Journal of Epidemiology 38*, 330–336.

Wakefield, J. C. (2004c). Ecological inference for 2×2 tables (with discussion). *Journal of the Royal Statistical Society, Series A 167*, 385–445.

Wakefield, J. C. (2004d). Response to the discuosion of "ecological inference for 2×2 tables". *Journal of the Royal Statistical Society, Series A 167*, 440–445.