Advance Access publication August 13, 2007

Political Analysis (2008) 16:41–69
doi:10.1093/pan/mpm017

Bayesian and Likelihood Inference for 2 × 2 Ecological Tables: An Incomplete-Data Approach

Kosuke Imai
Department of Politics, Princeton University, Princeton, NJ 08544
e-mail: kimai@princeton.edu (corresponding author)

Ying Lu
Department of Sociology, University of Colorado at Boulder, Boulder, CO 80309
e-mail: ying.lu@colorado.edu

Aaron Strauss
Department of Politics, Princeton University, Princeton, NJ 08544
e-mail: abstraus@princeton.edu

Ecological inference is a statistical problem where aggregate-level data are used to make inferences about individual-level behavior. In this article, we conduct a theoretical and empirical study of Bayesian and likelihood inference for 2 × 2 ecological tables by applying the general statistical framework of incomplete data. We first show that the ecological inference problem can be decomposed into three factors: *distributional effects*, which address the possible misspecification of parametric modeling assumptions about the unknown distribution of missing data; *contextual effects*, which represent the possible correlation between missing data and observed variables; and *aggregation effects*, which are directly related to the loss of information caused by data aggregation. We then examine how these three factors affect inference and offer new statistical methods to address each of them. To deal with distributional effects, we propose a nonparametric Bayesian model based on a Dirichlet process prior, which relaxes common parametric assumptions. We also identify the statistical adjustments necessary to account for contextual effects. Finally, although little can be done to cope with aggregation effects, we offer a method to quantify the magnitude of such effects in order to formally assess its severity. We use simulated and real data sets to empirically investigate the consequences of these three factors and to evaluate the performance of our proposed methods. C code, along with an easy-to-use R interface, is publicly available for implementing our proposed methods (Imai, Lu, and Strauss, forthcoming).

Authors' note: This article is in the part based on two working papers by Imai and Lu, "Parametric and Nonparamateric Bayesian Models for Ecological Inference in 2 × 2 Tables" and "Quantifying Missing Information in Ecological Inference." Various versions of these papers were presented at the 2004 Joint Statistical Meetings, the Second Cape Cod Monte Carlo Workshop, the 2004 Annual Political Methodology Summer Meeting, and the 2005 Annual Meeting of the American Political Science Association. We thank anonymous referees, Larry Bartels, Wendy Tam Cho, Jianqing Fan, Gary King, Xiao-Li Meng, Kevin Quinn, Phil Shively, David van Dyk, Jon Wakefield, and seminar participants at New York University (the Northeast Political Methodology conference), at Princeton University (Economics Department and Office of Population Research), and at the University of Virginia (Statistics Department) for helpful comments.

© The Author 2007. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

1 Introduction

Ecological inference is a statistical problem where aggregate-level data are used to make inferences about individual-level behavior. Although it was first studied by sociologists in the 1950s (Robinson 1950; Duncan and Davis 1953; Goodman 1953), recent years have witnessed resurgent interest in ecological inference among political methodologists and statisticians (see, e.g., Achen and Shively 1995; King 1997; King, Rosen, and Tanner 2004; Wakefield 2004a, and references therein). Much of the existing research, however, has focused on the development of new parametric models and the criticism of existing models and has generated numerous debates over the appropriateness of proposed methods and their use (see, e.g., Freedman et al. 1991; Grofman 1991; Cho 1998; Cho and Gaines 2004; Herron and Shotts 2004, and many others).

In this article, we conduct a theoretical and empirical study of Bayesian and likelihood inference for 2×2 ecological tables by applying the general statistical framework of incomplete (or missing) data (Heitjan and Rubin 1991).[1] First, we formulate ecological inference in 2×2 tables as a missing-data problem where only the weighted average of two unknown variables is observed (Section 2). This framework directly incorporates the deterministic bounds, which contain all information available from the data, and allows researchers to use the individual-level data whenever available. Within this general framework, we first show that the ecological inference problem can be decomposed into three factors: *distributional effects*, which address the possible misspecification of parametric modeling assumptions about the unknown distribution of missing data; *contextual effects*, which represent the possible correlation between missing data and observed variables; and *aggregation effects*, which are directly related to the loss of information caused by data aggregation.

We then examine how each of these three factors affects inference and offer new statistical methods to address each of them. To deal with distributional effects, we extend a simple parametric model to a nonparametric Bayesian model based on a Dirichlet process prior (Section 3). One common feature of many existing models is the use of parametric assumptions. In the exchange between King (1999) and Freedman et al. (1998), King concludes that "open issues . . . include . . . flexible distributional and functional form specifications" (354). We take up this challenge by relaxing the distributional assumption and examine the relative advantages of the proposed nonparametric model through simulation studies and an empirical example. We also show that statistical adjustments for contextual effects can be made within these parametric and nonparametric models.

Although little can be done to cope with aggregation effects, we offer a method to quantify the magnitude of such effects within our parametric model by quantifying the amount of missing information due to data aggregation in ecological inference (Section 4). Our approach is to measure the amount of information the observed aggregate-level data provide in comparison with the information one would obtain if the individual-level data were available. We do so in the context of both parameter estimation and hypothesis testing. Previous studies largely relied upon informal graphical and numerical summaries in order to examine the amount of information available in the observed data (e.g., King 1997; Gelman et al. 2001; Cho and Gaines 2004; Wakefield 2004a). In contrast, the proposed methods can be used to formally assess the severity of aggregation effects.

Finally, we evaluate the performance of our proposed methods and illustrate their use with the analysis of both simulated and real data sets (Section 5). C code, along with an

[1] See Cross and Manski (2002) and Judge, Miller, and Cho (2004) for alternative approaches, which are not based on the likelihood function.

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

Table 1 2 × 2 Ecological table for the racial voting example

	Black voters	*White voters*	
Voted	W_{i1}	W_{i2}	Y_i
Not voted	$1 - W_{i1}$	$1 - W_{i2}$	$1 - Y_i$
	X_i	$1 - X_i$	

Note. X_i, Y_i, W_{i1}, and W_{i2} are proportions and hence lie between 0 and 1. The unit of observation is typically a geographical unit and is denoted by i.

easy-to-use R interface, is publicly available as an R package, eco (Imai, Lu, and Strauss forthcoming), through the Comprehensive R Archive Network (http://cran.r-project.org/) for implementing our proposed methods.

2 Theoretical Framework for Ecological Inference

We first introduce a general theoretical framework for the ecological inference problem in 2 × 2 tables. We show that ecological data can be viewed as coarse data, which are a special case of incomplete data. Following the general framework of Heitjan and Rubin (1991), we discuss the conditions under which valid ecological inferences can be made using likelihood-based models. This theoretical framework clarifies and formally identifies the modeling assumptions required for ecological inference. While demonstrating how to deal with common problems, the framework also provides insight into the fundamental difficulty inherent in ecological inference, which cannot be overcome by statistical adjustments.

2.1 *Ecological Inference Problem in 2 × 2 Tables*

In this article, we focus on ecological inference in 2 × 2 tables. Suppose, for example, that we observe the number of registered white and black voters for each geographical unit (e.g., a county). The election results reveal the total number of votes for all geographical units. Given this information, we wish to infer the number of black and white voters who turned out. Table 1 presents this 2 × 2 ecological inference example, where counts are transformed into proportions. In typical political science examples, the number of voters within each geographical unit is large. Hence, many previous methods directly modeled proportions rather than counts (e.g., Goodman 1953; Freedman et al. 1991; King 1997).[2] We focus on models of proportions in this article.

For every geographical unit $i = 1, \ldots, n$, such a 2 × 2 ecological table is available. Given the total turnout rate Y_i and the proportion of black voters X_i, one seeks to infer the proportions of black and white voters who turned out, W_{i1} and W_{i2}, respectively. Although both W_{i1} and W_{i2} are not observed, they follow a key deterministic relationship,

$$Y_i = W_{i1}X_i + W_{i2}(1 - X_i). \tag{1}$$

That is, Y_i is the observed weighted average of the two unknown variables, W_{i1} and W_{i2}, with X_i and $1 - X_i$ being the observed weights.

The goals of ecological inference are twofold. First, researchers may be interested in characterizing the individual behavior at the population level. For example, they may wish to estimate the mean and variance of the joint or marginal (population) distributions of W_1 and W_2, or the distributions themselves. Second, since the internal cells of ecological tables are not observed, the estimation of the (sample) values of W_{i1} and W_{i2} for each unit

[2]See Brown and Payne (1986); King, Rosen, and Tanner (1999); and Wakefield (2004a) for models of counts.

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

i is also of interest. We call the former *population ecological inference*, whereas the latter is referred to as *sample ecological inference*. In political science research, sample ecological inference is often emphasized more often than population inference (e.g., King 1997). However, in other studies such as epidemiological studies that assess disease risk factors through ecological data, population ecological inference is of primary importance.

If sample ecological inference is conducted within the frequentist statistical framework, W_{i1} and W_{i2} should not be treated as unknown parameters to be estimated. In that case, we must estimate n parameters based on n observations, and no informational gain results from obtaining additional observations. Instead, each new observation creates an additional parameter to estimate. Such an approach yields an incidental parameter problem where no consistent estimators can be constructed for W_{i1} and W_{i2} (Neyman and Scott 1948). Hence, W_{i1} and W_{i2} must be viewed as missing data to be predicted rather than parameters to be estimated. The distinction between sample and population inferences, therefore, is critical for understanding the statistical properties of various frequentist ecological inference models.

2.2 *Ecological Inference as a Coarse Data Problem*

We now show that ecological inference in 2×2 tables can be viewed as a *coarse data* problem. Coarse data refer to a particular type of incomplete data that are neither entirely missing nor perfectly observed. Instead, we observe only a subset of the complete-data sample space in which the true unobserved data points lie. Some examples of coarse data include rounded, heaped, censored, and partially categorized data (Heitjan and Rubin 1991).

For ecological inference in 2×2 tables, the vector of internal cells $W_i = (W_{i1}, W_{i2})$ are the variables of interest. However, they are not directly observed. Instead, only their weighted average Y_i and the weight X_i are observed. From equation (1), Duncan and Davis (1953) derive the sharp bounds for each of the unobserved variables, W_{i1} and W_{i2},

$$\begin{aligned} W_{i1} &\in \left[\max\left(0, \frac{X_i + Y_i - 1}{X_i}\right), \min\left(1, \frac{Y_i}{X_i}\right)\right], \\ W_{i2} &\in \left[\max\left(0, \frac{Y_i - X_i}{1 - X_i}\right), \min\left(1, \frac{Y_i}{1 - X_i}\right)\right]. \end{aligned} \tag{2}$$

Although these intervals reveal the possible values that W_{i1} and W_{i2} could take, they are often too wide to be informative for the purposes of applied researchers.

Ecological inference is a coarse data problem because the missing data $W_i = (W_{i1}, W_{i2})$ are only partially observed. The relationship between the observed data (Y_i, X_i) and the missing data W_i is solely characterized by equation (1). The random variable X_i is called a coarsening variable, whereas Y_i is called the coarsened data. This terminology is derived from the fact that X_i determines how much information is revealed about each of the missing data through Y_i. For example, if there are many more black voters than white voters, then the aggregate turnout rate gives you more information about black voters' turnout. In other words, if X_i takes a value closer to 1, bounds are likely to be narrow for W_{i1} and wide for W_{i2}.

2.3 *Three Key Factors in Ecological Inference*

Next, we place ecological inference within the theoretical framework of coarse data developed by Heitjan and Rubin (1991) and formally identify the key factors that influence ecological inference. We consider the likelihood-based inference, which has been a popular approach in the literature (e.g., King 1997; King, Rosen, and Tanner 1999; Rosen et al.

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

2001; Wakefield 2004a). We begin by defining the many-to-one mapping, $Y_i = \mathcal{M}(X_i, W_i) = X_i W_{i1} + (1 - X_i)W_{i2}$, from the complete data to the observed (coarsened) data for each $i = 1, 2, \ldots, n$. Suppose that the density function of W_i is given by $f(W_i \mid \zeta)$ with a vector of unknown parameters ζ. Let $h(X_i \mid W_i, \gamma)$ denote the conditional distribution of X_i given unobserved data W_i and a vector of unknown parameters γ. Then, the observed-data likelihood function can be written as,

$$L_{\text{obs}}(\zeta, \gamma \mid Y, X) = \prod_{i=1}^{n} \int_{Y_i = \mathcal{M}(X_i, W_i)} h(X_i \mid W_i, \gamma) f(W_i \mid \zeta) \mathrm{d}W_i, \tag{3}$$

where ζ and γ are assumed to be disjoint sets of parameters. The calculation of the observed-data likelihood function in equation (3) requires the integration with respect to the missing data W_i over the region defined by the data coarsening mechanism, $Y_i = \mathcal{M}(X_i, W_i)$. In contrast, the complete-data likelihood function, that is, the likelihood function one would obtain if the missing data were to be completely observed, is given by,

$$L_{\text{com}}(\zeta, \gamma \mid W, X) = \prod_{i=1}^{n} h(X_i \mid W_i, \gamma) f(W_i \mid \zeta). \tag{4}$$

To make inferences based on $L_{\text{obs}}(\zeta, \gamma \mid Y, X)$, we must specify the sampling distribution of missing data $f(W_i \mid \zeta)$ as well as the conditional distribution of the coarsening variable $h(X_i \mid W_i, \gamma)$. In ecological inference, this incomplete-data framework allows us to formally identify the following three key factors. The first factor is distributional effects, which refer to the effects of the (mis)specification of $f(W_i \mid \zeta)$ or the joint distribution of black and white turnout rates in our running example, on the resulting inference. The second factor is contextual effects, which are concerned about the specification of $h(X_i \mid W_i, \gamma)$. In our running example, the proportion of black voters might be correlated with black and white turnout rates through neighborhood variables such as income and education. The debate in the literature has almost exclusively focused on the possible misspecification of these two distributions. Unfortunately, since W_i is not directly observed, detecting distributional and contextual effects is a difficult task in practice. For example, one can compare the marginal distribution of Y against the (marginal) predictive distribution of Y from the fitted model. Such an approach, however, will not be able to detect all the misspecified models because the misspecification of the distribution of W_i can still yield the marginal predictive distribution of Y that is consistent with the observed data. In Section 4.3, we partially address this concern of undetectable model misspecification under parametric assumptions.

Finally, we also study the third, yet most critical, issue of ecological inference, that is, the loss of information that occurs due to data coarsening. We call this aggregation effects because it is the data aggregation that makes ecological inference a fundamentally difficult statistical problem. Aggregation effects cause both distributional and contextual effects because the data aggregation prevents researchers from detecting model misspecification through the diagnostic techniques available to usual analysis of complete data. Although aggregation effects cannot be overcome by statistical adjustments, we show that it is possible to quantify the amount of missing information due to aggregation in ecological inference (Section 4).

2.4 *Three Modeling Assumptions*

Based on the theoretical framework introduced above, we identify three possible modeling assumptions for ecological inference and derive the general conditions under which valid ecological inferences can be drawn.

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

2.4.1 Assuming no contextual effect

First, we state the condition under which the stochastic coarsening mechanism can be ignored; that is, the condition under which the specification of $h(X_i \mid W_i, \gamma)$ is not required. In ecological inference, this corresponds to the condition under which contextual effects can be ignored. In our running example, this means that black and white turnout rates are jointly independent of the proportion of black voters. Although this is a strong assumption and often cannot be justified in practice, it serves as a useful starting point for developing models under more general conditions. Heitjan and Rubin (1991) formally define this condition and call it *coarsened at random* (CAR) as a general formulation of missing at random in the literature on inference with missing data.

Under CAR, if ζ and γ are disjoint parameters, the inference about ζ does not depend on γ and the specification of $h(X_i \mid W_i, \gamma)$ can be ignored. Heitjan and Rubin (1991) also show that CAR is the weakest condition under which it is appropriate to ignore the coarsening mechanism. Formally, a sufficient condition for Y_i to be CAR is that X_i and W_i are independent; that is, $h(X_i \mid W_i, \gamma) = h(X_i \mid \gamma)$. Then, the observed-data likelihood function of equation (3) can be simplified as

$$L_{\text{obs}}(\zeta \mid Y, X) = \prod_{i=1}^{n} \int_{Y_i = \mathcal{M}(X_i, W_i)} f(W_i \mid \zeta) \mathrm{d}W_i.$$

Parametric models under this assumption have appeared in the literature (e.g., King 1997; Wakefield 2004a).

2.4.2 Modeling contextual effects with covariates

In many situations, W_i and X_i may not be independent, but this dependence can be modeled through controlling for observed covariates Z_i, which may or may not include X_i. Another motivation for this approach is the estimation of the conditional mean function of W_i given Z_i rather than its marginal mean. We refer to this modeling assumption as *conditionally coarsened at random*, or CCAR. In the context of our running example, one may assume that once we control for income and education levels, black and white turnout rates are no longer dependent on the proportion of black voters.

Formally, we assume that W_i and X_i are conditionally independent given Z_i, that is, $h(X_i \mid W_i, Z_i, \gamma) = h(X_i \mid Z_i, \gamma)$. If the assumption holds, the data are CAR given Z_i, and the observed-data likelihood can be written as

$$L_{\text{obs}}(\zeta \mid Y, X, Z) = \prod_{i=1}^{n} \int_{Y_i = \mathcal{M}(X_i, W_i)} f(W_i \mid Z_i, \zeta) \mathrm{d}W_i.$$

King (1997) and King, Rosen, and Tanner (1999) propose parametric models based on this assumption.

2.4.3 Modeling contextual effects without covariates

Finally, we consider a scenario where the CAR assumption is known to be violated but no covariate is available for which the CCAR assumption holds. Even when some covariates are available, researchers may not be willing to make functional-form assumptions about the high-dimensional covariate space because we do not directly observe W_i. Unless we

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

jointly observe (W_i, X_i, Z_i) for some units, the *not coarsened at random* (NCAR) strategy is to minimize the modeling assumptions by focusing on the trivariate relationship between W_i and X_i without incorporating Z_i. In addition, one may wish to focus on the estimation of marginal mean of W_i rather than its conditional mean. We refer to this modeling assumption as NCAR. In the NCAR case, we directly model the data coarsening mechanism and specify the joint distribution $g(X_i, W_i \mid \zeta, \gamma) = f(W_i \mid \zeta)h(X_i \mid W_i, \gamma)$. The observed-data likelihood can be written as,

$$L_{\text{obs}}(\zeta, \gamma \mid Y, X) = \prod_{i=1}^{n} \int_{Y_i = \mathcal{M}(X_i, W_i)} g(X_i, W_i \mid \zeta, \gamma) \mathrm{d}W_i.$$

To the best of our knowledge, no model under this assumption has been proposed in the literature.

3 A Nonparametric Model of Ecological Inference

In this section, we introduce a Bayesian, nonparametric model of ecological inference in order to deal with distributional effects (as well as contextual effects) by relaxing parametric assumptions. We start our discussion by describing a parametric model, which is similar to the ones proposed in the literature, and then show how to extend the model to a nonparametric model.

3.1 *A Parametric Base Model*

Our first parametric model is based on the CAR assumption. A similar parametric model has appeared in the literature (King 1997; Wakefield 2004a). In particular, we model the logit transformation of the missing data using the bivariate normal distribution,

$$W_i^* \mid \mu, \Sigma \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma),$$

where $W_i^* = (W_{i1}^*, W_{i2}^*) = (\text{logit}(W_{i1}), \text{logit}(W_{i2}))$, μ is a (2×1) vector of population means, and Σ is a (2×2) positive-definite variance matrix. The model allows W_{i1} and W_{i2} to be correlated with each other (through their logit transformations). This means that in the racial voting example, the turnout rates of black and white voters in each county may be correlated with one another.

The above model can be extended to a Bayesian model by placing the following conjugate prior distribution on (μ, Σ),

$$\mu \mid \Sigma \sim \mathcal{N}\left(\mu_0, \frac{\Sigma}{\tau_0^2}\right), \quad \text{and} \quad \Sigma \sim \text{InvWish}(\nu_0, S_0^{-1}), \tag{5}$$

where μ_0 is a (2×1) vector of the prior mean, τ_0 is a scalar, ν_0 is the prior degrees of freedom parameter, and S_0 is a (2×2) positive-definite prior scale matrix. When strong prior information is available from previous studies or elsewhere, we specify these prior parameters so that the prior knowledge can be approximated. When such information is not available, however, we consider a flat prior where the prior predictive distribution of (W_1, W_2) is approximately uniform. This latter condition leads to our choice of the prior parameters for the parametric model: $\mu_0 = \mathbf{0}$, $S_0 = 10I_2$, $\tau_0 = 2$, and $\nu_0 = 4$.

This parametric base model can be easily extended to the analyses under the CCAR and NCAR assumptions. For example, under the CCAR assumption, the model becomes,

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

$$W_i^* \mid \beta, \Sigma, Z_i \overset{\text{indep.}}{\sim} \mathcal{N}(Z_i^{\mathrm{T}}\beta, \Sigma),$$

where β is a $(k \times 1)$ vector of coefficients, Z_i is a $(k \times 2)$ matrix of covariates, and Σ is the (2×2) positive-definite conditional variance matrix. In contrast, under NCAR, the model is specified as,

$$(W_i^*, X_i^*) \mid \eta, \Phi \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\eta, \Phi),$$

where $X_i^* = \text{logit}(X_i)$, the mean vector η is 3×1, and the covariance matrix Φ is a 3×3 positive-definite matrix.

3.2 *A Nonparametric Model*

Similar to other parametric models in the literature, the models introduced in Section 3.1 make specific distributional assumptions. To relax these assumptions, we apply a Dirichlet process prior and model the unknown population distribution as a mixture of bivariate normal distributions (Ferguson 1973).[3] The resulting model is nonparametric in the sense that no distributional assumption is made, and its in-sample predictions respect the deterministic bounds. Recent development of Markov chain Monte Carlo (MCMC) algorithms has enabled the use of a Dirichlet process prior for Bayesian density estimation and other nonparametric and semiparametric problems (e.g., Escobar and West 1995; Mukhopadhyay and Gelfand 1997; Gill and Casella 2006). Dey, Müller, and Sinha (1998) is an accessible introduction to this methodology.

Our basic idea is to use the (countably infinite) mixture of bivariate normal distributions to model the unknown distribution of W. Unlike finite mixture models, the number of mixtures (or clusters) is not specified in advance and can grow as the number of data points increases, thereby allowing for nonparametric estimation of an unknown distribution. In fact, each new draw of the data may come from one of the existing mixture components from which the other data points were generated or from a new distribution adding another component to the mixture. The number of mixture components is controlled by a single parameter, α, which is a positive scalar and called the concentration parameter. Our model specifies a prior distribution on α which results in a relatively large number of mixture components, and then through posterior updating we learn about the number of clusters from the observed data.

Formally, we model the parameters, $\{\mu_i, \Sigma_i\}_{i=1}^n$, with an unknown (random) distribution function G rather than a known (fixed) one such as the normal/inverse-Wishart distribution. Note that the parameters now have subscript i, allowing for the possibility that the number of parameters grows as the number of observation grows (i.e., nonparametric estimation). We then place a prior distribution on G over all possible probability measures. Such a prior distribution is called a Dirichlet process prior and is denoted by $G \sim \mathcal{D}(G_0, \alpha)$, where $G_0(\cdot)$ is the known base prior distribution and is also the prior expectation of $G(\cdot)$; $E(G(\mu, \Sigma)) = G_0(\mu, \Sigma)$ for all (μ, Σ) in its parameter space. Ferguson (1973) established that given any measurable partition $(A_1, A_2, \ldots, A_k)$ on the support of G_0, the random vector of probabilities $(G(A_1), G(A_2), \ldots, G(A_k))$ follows a Dirichlet distribution with parameter $(\alpha G_0(A_1), \alpha G_0(A_2), \ldots, \alpha G_0(A_k))$. A large value of α suggests that G is likely to be close to G_0 and, hence, to yield the results that are similar to those obtained from the parametric model with the prior distribution G_0. On the other hand, a small value of α implies that G is likely to place most of the probability mass on a few

[3] See Imai and King (2004) for an alternative approach based on the Bayesian model averaging.

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

partitions. This setup allows the unknown distribution function G to be nonparametrically estimated from the data.

We specify a Dirichlet process prior on the unknown distribution function of the population parameters, using the same conjugate normal/inverse-Wishart prior distribution as the base prior distribution. Finally, we place a gamma prior on the concentration parameter α. Then, our Bayesian nonparametric model is given by,

$$\begin{aligned} W_i^* \mid \mu_i, \Sigma_i &\sim \mathcal{N}(\mu_i, \Sigma_i), \\ \mu_i, \Sigma_i \mid G &\sim G, \\ G \mid \alpha &\sim \mathcal{D}(G_0, \alpha), \\ \alpha &\sim \mathcal{G}(a_0, b_0), \end{aligned}$$

where under G_0, (μ_i, Σ_i) is distributed as

$$\mu_i \mid \Sigma_i \sim \mathcal{N}\left(\mu_0, \frac{\Sigma_i}{\tau_0^2}\right), \quad \text{and} \quad \Sigma_i \sim \text{InvWish}(\nu_0, S_0^{-1}).$$

To illustrate how our model relates to a normal mixture, we follow Ferguson (1973) and Escobar and West (1995) to compute the conditional prior distribution, $p(\mu_i, \Sigma_i \mid \mu^{(i)}, \Sigma^{(i)}, \alpha)$, where $\mu^{(i)} = \{\mu_1, \ldots, \mu_{i-1}, \mu_{i+1}, \ldots, \mu_n\}$ and $\Sigma^{(i)} = \{\Sigma_1, \ldots, \Sigma_{i-1}, \Sigma_{i+1}, \ldots, \Sigma_n\}$. The calculation yields,

$$\mu_i, \Sigma_i \mid \mu^{(i)}, \Sigma^{(i)}, \alpha \sim \alpha\, a_{n-1} G_0(\mu_i, \Sigma_i) + a_{n-1} \sum_{j=1,\, j \neq i}^{n} \delta_{(\mu_j, \Sigma_j)}(\mu_i, \Sigma_i) \quad \text{for } i = 1, \ldots, n, \tag{6}$$

where $\delta_{(\mu_j, \Sigma_j)}(\mu_i, \Sigma_i)$ is a degenerate distribution whose entire probability mass is concentrated at $(\mu_i, \Sigma_i) = (\mu_j, \Sigma_j)$ and $a_{n-1} = 1/(\alpha + n - 1)$. Equation (6) shows that given any $(n - 1)$ values of (μ_i, Σ_i), there is a positive probability of coincident values and that as α tends to ∞, the distribution approaches G_0. In other words, a new draw of the parameters can take either the same values as one of the existing parameter values or new values drawn from the base distribution. The relative frequencies of these two events are governed by the concentration parameter α.

Similarly, a future replication draw of $(\mu_{n+1}, \Sigma_{n+1})$, given $\mu = \{\mu_1, \ldots, \mu_n\}$ and $\Sigma = \{\Sigma_1, \ldots, \Sigma_n\}$, has the mixture distribution,

$$\mu_{n+1}, \Sigma_{n+1} \mid \mu, \Sigma, \alpha \sim \alpha\, a_n G_0(\mu_{n+1}, \Sigma_{n+1}) + a_n \sum_{i=1}^{n} \delta_{(\mu_i, \Sigma_i)}(\mu_{n+1}, \Sigma_{n+1}),$$

where $a_n = 1/(\alpha + n)$. We then compute the predictive distribution of a future observation W_{n+1}^* given (μ, Σ, α), which forms the basis of Bayesian density estimation. In particular, we evaluate $\int p(W_{n+1}^* \mid \mu_{n+1}, \Sigma_{n+1}, \alpha) \mathrm{d}P(\mu_{n+1}, \Sigma_{n+1} \mid \mu, \Sigma, \alpha)$, which yields,

$$W_{n+1}^* \mid \mu, \Sigma, \alpha \sim \alpha\, a_n \mathcal{T}_{\nu_0}(\mu_0, S) + a_n \sum_{i=1}^{n} \mathcal{N}(\mu_i, \Sigma_i), \tag{7}$$

where $\mathcal{T}_{\nu_0}(\mu_0, S)$ is a bivariate t distribution with ν_0 degrees of freedom, the location parameter μ_0, and the scale matrix $S = (\tau_0^2 + 1)S_0/\{\tau_0^2(1 + \nu_0)\}$. Equation (7) shows that when the value of α is small, the predictive distribution is equivalent to a normal mixture. This setup resembles the standard kernel density estimator with a bivariate normal kernel.

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

In particular, α plays a role similar to the bandwidth parameter, which controls the degree of smoothness.

We use a diffuse prior, $\mathcal{G}(1, 0.1)$, with a mean of 10 and variance 100 for the concentration parameter, α. According to Antoniak (1974), the expected number of clusters given α and the sample size n is approximately $\alpha \log(1 + n/\alpha)$. With this choice of prior distribution for α and $n = 200$, the prior expected number of clusters is approximately 27. In general, a sensitivity analysis should be conducted in order to assess the influence of prior specification on posterior inferences. The sensitivity analysis is important especially for the concentration parameter because it plays a critical role in the density estimation with Dirichlet processes.

This nonparametric model can be easily extended to the analysis under the NCAR assumption by placing the following conjugate prior distribution on (η, Φ); that is, $\eta \mid \Phi \sim \mathcal{N}(\eta_0, \Phi/\tau_0^2)$ and $\Phi \sim \text{InvWish}(\nu_0, S_0^{-1})$, where ν_0 is the (3×1) vector of prior mean, $\tau_0 > 0$ is a scale parameter, ν_0 is the prior degrees of freedom parameter, and S_0 is the (3×3) positive-definite prior scale matrix. For the inverse-Wishart distribution to be proper, ν_0 needs to be greater than 3.

Our nonparametric model, therefore, in principle can provide flexible estimation of bivariate density functions for ecological inference problems. However, because we do not directly observe W_{i1} and W_{i2}, the density estimation problem for ecological inference is much more difficult. Therefore, bounds must be sufficiently informative in order for the nonparametric model to be able to recover the underlying population distribution. We empirically investigate this issue through both the analysis of simulated and real data sets in Section 5.1.

3.3 *Computational Strategies*

Finally, we briefly discuss our computational strategies to fit the proposed models. To obtain the maximum likelihood (ML) estimates of the model parameters for the parametric CAR and NCAR models, we develop an Expectation Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977), whereas we develop an Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin 1993) to fit the CCAR model. The details of these algorithms appear in Appendix A. The EM and ECM algorithms are general optimization techniques that are often useful when obtaining the ML estimates in the presence of missing data. A main advantage of these algorithms is their numerical stability. In particular, the observed-data log likelihood increases monotonically at each iteration. For Bayesian analysis, we develop MCMC algorithms for both parametric and nonparametric models. These MCMC algorithms are described in Appendix B.

4 Quantifying the Aggregation Effects

In this section, under the theoretical framework described in Section 2, we show how to quantify the magnitude of the aggregation effects in the context of both parameter estimation and hypothesis testing. We do so by measuring the fraction of missing information under the parametric models proposed in Section 3.1. Our approach is to quantify the amount of missing information caused by data aggregation relative to the amount of information one would have if the individual-level data are observed.

4.1 *A Measure of the Aggregation Effects in Parameter Estimation*

To quantify the amount of the aggregation effects in parameter estimation, we use the missing-information principle of Orchard and Woodbury (1972), which states that

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

the *missing information* is equal to the difference between the *complete information* and the *observed information*. Formally, Dempster, Laird, and Rubin (1977) prove the following key equality,

$$\mathcal{I}_{\text{mis}}(\hat{\theta}) = \mathcal{I}_{\text{com}}(\hat{\theta}) - \mathcal{I}_{\text{obs}}(\hat{\theta}),$$

where $\hat{\theta}$ is the ML estimate of unknown parameters θ ($\theta = (\xi, \gamma)$ in our case) from the observed data. $\mathcal{I}_{\text{obs}}(\hat{\theta})$ represents the observed Fisher information matrix, defined by,

$$\mathcal{I}_{\text{obs}}(\hat{\theta}) \equiv -\frac{\partial^2}{\partial\theta^2} l_{\text{obs}}(\theta \mid Y, X)\Big|_{\theta=\hat{\theta}}, \tag{8}$$

where l_{obs} is the observed-data log-likelihood function based on equation (3). $\mathcal{I}_{\text{com}}(\hat{\theta})$ denotes the expected information matrix from the complete-data log-likelihood function, based on equation (4), and is given by,

$$\mathcal{I}_{\text{com}}(\hat{\theta}) \equiv E\left[-\frac{\partial^2}{\partial\theta^2} l_{\text{com}}(\theta \mid W, X)\Big| Y, X, \theta\right]\Big|_{\theta=\hat{\theta}}, \tag{9}$$

where the expectation is taken with respect to the distribution of missing data W given the observed data (Y, X). Finally, $\mathcal{I}_{\text{mis}}(\hat{\theta})$ can be viewed as the missing information due to data aggregation and is defined as,

$$\mathcal{I}_{\text{mis}}(\hat{\theta}) \equiv E\left[-\frac{\partial^2}{\partial\theta^2} \log p(W \mid X, Y, \theta)\Big| Y, X, \theta\right]\Big|_{\theta=\hat{\theta}}.$$

To define a measure of missing information in multivariate settings, we use the diagonal elements of the (matrix) fraction of the observed information and complete information,

$$F_\theta \equiv \text{diag}(I - \mathcal{I}_{\text{obs}}(\hat{\theta})\mathcal{I}_{\text{com}}(\hat{\theta})^{-1}) \tag{10}$$

Then, the ith element of F_θ is an information-theoretic measure of the relative amount of missing information in the ML estimation of the ith element of the parameter vector θ. In ecological inference, F_θ represents the amount of additional information the individual-level data would provide for the estimation of θ, if they were available, in comparison with the information obtained from the observed aggregate data. Since the diagonal elements of the inverse of the observed information matrix equal the estimated asymptotic variance of each parameter, in the one-parameter case, the fraction of missing information equals the fraction of increase in the asymptotic variance due to missing data.

Finally, it is also possible to summarize the amount of missing information in ecological inference by a scalar rather than computing the fraction of missing information for each parameter. This can be done by computing the largest eigenvalue of the "matrix fraction" of missing information, $I - \mathcal{I}_{\text{com}}^{-1}(\hat{\theta})\mathcal{I}_{\text{obs}}(\hat{\theta})$, where I represents the identity matrix. In this expression, a larger value indicates a greater amount of missing information.

4.2 *A Measure of the Aggregation Effects in Hypothesis Testing*

Kong, Meng, and Nicolae (2005) propose a general framework for quantifying the relative amount of missing information in hypothesis testing with incomplete data. We apply this methodology to ecological inference so that the fraction of missing information can be calculated for hypothesis testing. Kong, Meng, and Nicolae (2005) propose two measures of missing information in hypothesis testing: the fraction of missing information against

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

and under a null hypothesis. In this article, we focus on the former because, as discussed by Kong, Meng, and Nicolae (2005), the latter may provide misleading inferences if the true values are far away from the null values.

Consider the null hypothesis $H_0 : \theta = \theta_0$. The fraction of missing information against the null hypothesis is given by,

$$F_H \equiv 1 - \frac{l_{\text{obs}}(\hat{\theta} \mid Y, X) - l_{\text{obs}}(\theta_0 \mid Y, X)}{E[l_{\text{com}}(\hat{\theta} \mid W, X) - l_{\text{com}}(\theta_0 \mid W, X) \mid Y, X; \hat{\theta}]}, \tag{11}$$

where the expectation is taken over the conditional distribution of the missing data W given the observed information (Y, X). This measure equals the ratio of the logarithms of the two likelihood ratio test statistics; the logarithm of *the observed likelihood ratio statistic*, based on the observed-data likelihood, is in the numerator whereas the logarithm of *the expected likelihood ratio statistic*, based on the complete-data likelihood, is in the denominator.

The interpretation of the measure in equation (11) exactly parallels that of the fraction of missing information in parameter estimation (see equation 10). Kong, Meng, and Nicolae (2005) show the three key properties of this measure; (1) F_H is a fraction, that is, $0 \leq F_H \leq 1$; (2) $F_H = 1$ if and only if the observed data cannot distinguish between $\hat{\theta}$ and θ_0 at all; that is, the observed-data likelihood ratio is equal to 1 or $l_{\text{obs}}(\hat{\theta} \mid Y, X) = l_{\text{obs}}(\theta_0 \mid Y, X)$; and (3) $F_H = 0$ if and only if the missing information cannot distinguish between $\hat{\theta}$ and θ_0 given the observed data; that is, the Kullback-Leibler information number, $E\left[\log \frac{p(W \mid Y, X, \hat{\theta})}{p(W \mid Y, X, \theta_0)} \mid Y, X; \hat{\theta}\right]$, is equal to 0.

4.2.1 Null hypothesis of linear constraints on marginal means

We first consider the null hypothesis of linear constraints on the marginal means of W_i under the CAR and NCAR models. If we have l linear constraints, then the null hypothesis can be written as the system of l linear equations, $H_0 : \mathbf{A}^{\mathrm{T}}\mu = a$, where a is an $(l \times 1)$ vector of known constants. For the CAR model, μ is a two-dimensional vector, whereas under the NCAR model it is a three-dimensional vector. An important special case is the equality constraint of marginal means, $\mu_1 = \mu_2$ or equivalently $\mathbf{A} = (1, -1)$ and $a = 0$ under the CAR model and $\mathbf{A} = (1, -1, 0)$ and $a = 0$ under the NCAR model. For example, researchers may wish to test whether the turnout rates of whites and nonwhites are the same. To conduct the likelihood ratio test of H_0 and compute the fraction of missing information associated with it, we must first obtain the ML estimates of θ under the constraint of $\mathbf{A}^{\mathrm{T}}\mu = a$, and then compare the value of the observed-data log likelihood under this constraint with the corresponding value obtained without the constraint.

4.2.2 Null hypothesis of linear constraints on regression coefficients

We next consider the null hypothesis of linear constraints on regression coefficients under the CCAR model. For example, one might be interested in testing the null hypothesis that the effect of a particular variable is zero on the conditional means of both W_{i1}^* and W_{i2}^*. If there are l linear constraints, the null hypothesis can be expressed as a system of l linear equations, $H_0 : A^{\mathrm{T}}\beta = a$ where $\mathbf{A}$ is a known $(k \times l)$ matrix and a is an l-dimensional vector of constants.

4.3 *Missing Information and Model Misspecification*

The proposed methods to quantify the amount of missing information described above assume that researchers know the correct (likelihood-based) parametric model. Although

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

most social scientists conduct their data analysis based on such an assumption, the possibility of model misspecification is greater in ecological inference and hence this is a potential concern. Given that the individual-level data are partially missing, standard diagnostics tools, which require complete data, cannot be used to detect possible model misspecification. This means that if the underlying complete-data model is incorrect, the resulting estimates of fraction of missing information may also be misleading. This problem reflects the fundamental difficulty of statistical inference in the presence of missing data; the inference may be sensitive to the modeling assumptions about missing data. Therefore, it is no surprise that much methodological controversy in ecological inference is centered around the issue of model misspecification.

Methodological research has only begun to directly address the problem of model uncertainty in ecological inference (e.g., Imai and King 2004). The methods we propose in this article, however, have only an indirect relationship with the issue of model misspecification. Namely, a higher fraction of missing information implies a greater magnitude of possible incomplete-data bias resulting from *local model misspecification*, that is, the degree of model misspecification which cannot be detected even if one would observe the complete data. Copas and Eguchi (2005) formalize this idea by showing that the magnitude of standardized incomplete-data bias for parameter θ resulting from such local model misspecification has the upper bound, which is equal to $\varepsilon\sqrt{F_\theta}$, where ε represents the magnitude of local model misspecification, and F_θ is the fraction of missing information in the estimation of θ as defined in equation (10). This formulation implies that the methods proposed in this section can alert applied researchers to the possibility of local model misspecification. However, the fraction of missing information does not reflect the degree to which the assumed model is grossly misspecified.

In ecological inference, such undetectable, yet serious, model misspecification might occur so that additional aggregate data (or coarse data) do not help detect the misspecification of individual-level data (or complete data) model. In that case, the magnitude of bias is likely to be larger than the above upper bound, and hence, the fraction of missing information may even underestimate the degree of model uncertainty.

4.4 *Computational Strategies*

To compute a measure of missing information under the CAR model, we apply the supplemented EM (SEM) algorithm to compute the fraction of missing information defined in equation (10) and to estimate the asymptotic variance-covariance matrix of the ML estimates (Meng and Rubin 1991). In addition to its numerical stability, a principle advantage of the SEM algorithm is that it simply extends the EM process, obviating the need to develop an independent algorithm.

Since the EM algorithm outputs the ML estimates of transformed parameters, $\theta^* = (\mu_1, \mu_2, \log \sigma_1, \log \sigma_2, 0.5 \log [(1 + \rho)/(1 - \rho)])$, we first compute the fraction of missing information for each of the transformed parameters. We use $\hat{\theta}^*$ to denote the ML estimates of transformed parameters θ^*. It is also possible to present the fraction of missing information for the parameters that can be easily interpreted by applied researchers rather than the transformed parameters, θ^*, which are used purely for the modeling and computational purposes. In this case, we use the first-order approximation to calculate the means, variances, and correlation of the original data, for example, W_{ij}, by $\text{logit}^{-1}(\mu_j)$, $\sigma_j e^{2\mu_j}/(1 + e^{\mu_j})^4$, and ρ, respectively. We then use the chain rule and the invariance property of ML estimators to derive the expression for the *DM* matrix and the expected information matrix, $\mathcal{I}_{\text{com}}$, for the new parameters of interest. A similar estimation strategy can be used for the NCAR model as well.

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

For the CCAR model, we use the supplemented ECM (SECM) algorithm, which modifies the SEM algorithm to adjust for the fact that the conditional maximization is used (van Dyk, Meng, and Rubin 1995). Using the SECM algorithm, we first compute the fraction of missing information for the transformed parameters, θ^*. If desired, we can also compute the fraction of missing information on the conditional mean on the logit scale, that is, $E(W_i^* \mid Z_i)$, or on the original scale, that is, $E(W_i \mid Z_i)$, using the first-order approximation and the invariance property of ML estimators.

5 Simulation Studies and Empirical Examples

In this section, we evaluate the performance of the proposed methods and illustrate their use by analyzing both simulated and real data sets. Each of the following subsections focuses on one of the three key factors in ecological inference identified in Section 2.

5.1 *Distributional Effects*

5.1.1 A simulation study

To investigate distributional effects, we use X_i from the data set analyzed by Burden and Kimball (1998), which has a sample size of 361. Although this data set is not about racial voting, for simplicity, we use the notation of Table 1 and refer to X_i as the proportion of black voters and Y_i as the overall turnout rate for each county i. The unknown inner cells (W_{i1}, W_{i2}) are the fractions of those who voted among black and white voters, respectively. To construct different simulation settings, we draw (W_{i1}, W_{i2}) independently from the following three distributions, although maintaining the same racial composition X_i.

Simulation I. W_i^* is independently drawn from a bivariate normal distribution with mean (0, 1.4), variances (1, 0.5), and covariance 0.2, yielding the average turnout of about 50% and 80% for black and white voters, respectively.

Simulation II. W_i^* is independently drawn from a mixture of two bivariate normal distributions with the mixing probability (0.6, 0.4). The first distribution has mean $(-0.4, 1.4)$, variance $(0.2, 0.1)$, and covariance 0. The second distribution has a different mean $(-0.4, -1.4)$, but the same covariance matrix. This yields the average turnout of roughly 40% for black voters, approximately 80% for white voters in 60% of the counties, and about 20% for white voters in the other counties.

Simulation III. W_i^* is independently drawn from a mixture of two bivariate normal distributions with the mixing probability (0.6, 0.4). The first distribution has mean $(-1.4, 1.4)$, variance $(0.1, 0.1)$, and covariance 0. The second distribution has a different mean $(1.4, -1.4)$, but the same covariance matrix. In 60% of the counties, the average turnout is 20% for blacks and 80% for whites, whereas in the rest of the counties this pattern is reversed.

In all three simulations, we assume no contextual effect. Note that in Simulation II only the marginal distribution of W_{i2} is bimodal, whereas in Simulation III the marginal distributions of both W_{i1} and W_{i2} are bimodal. It is of particular interest to see whether the nonparametric method can recover such distributions.

Figure 1 presents the tomography plots of the simulated data sets with the true values of W_i. The graphs illustrate the bounds for W_{i1} and W_{i2}, which can be obtained by projecting tomography lines onto the horizontal and vertical axes. The average length of bounds for W_{i1} in Simulations I, II, and III is 0.55, 0.58, and 0.64, whereas that for W_{i2} is 0.71, 0.73,

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

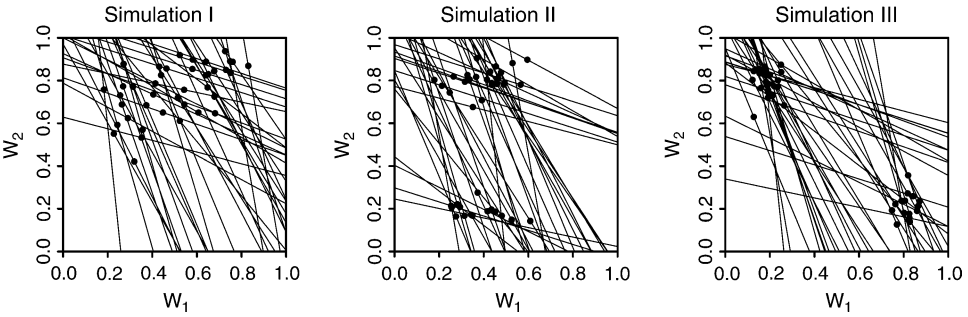


Fig. 1 Tomography plots of simulations I, II, and III. The solid lines illustrate the deterministic relationship of equation (2), and the dots represent the true values of (W_{i1}, W_{i2}), for randomly selected 40 counties from the Burden and Kimball (1998) data set.

and 0.78, respectively. This indicates that in all three simulations, the bounds are not particularly informative.

Treating X_i and Y_i as observed and W_i as unknown, we fit our parametric and nonparametric models and assess their relative performance in terms of both sample and population inferences by examining in-sample and out-of-sample predictions, respectively. Table 2 numerically summarizes the in-sample predictive performance. In Simulations II and III, the (sample) root mean squared error (RMSE) of our nonparametric model is smaller than that of the parametric model. Nevertheless, even when the true distribution is bimodal, the in-sample predictions from our parametric model are reasonable. This is because the parametric model yields the in-sample predictions that respect the bound conditions. The in-sample predictions based on the ecological regression (Goodman 1953) $E(Y_i \mid X_i) = \alpha + \beta X_i$ yield larger bias and RMSE than the other two methods.

Finally, we examine the out-of-sample predictive performance, which is of importance for population inferences. Figure 2 compares the true distribution with the estimated marginal density based on out-of-sample predictions from our models. In Simulation I, our nonparametric and parametric models give essentially identical estimates and approximate the marginal distributions well. Indeed, the number of clusters for the nonparametric model reduces to one. In our setup, the nonparametric model with one cluster is identical

Table 2 In-sample predictive performance with different distributions of (W_1, W_2)

	Simulation I		*Simulation II*		*Simulation III*	
	W_1	W_2	W_1	W_2	W_1	W_2
Bias						
Parametric model	−0.004	0.001	0.003	−0.012	−0.010	0.010
Nonparametric model	−0.004	0.001	0.008	−0.011	−0.010	0.009
Ecological regression	−0.011	0.011	−0.003	0.006	−0.027	0.029
RMSE						
Parametric model	0.084	0.080	0.098	0.166	0.134	0.137
Nonparametric model	0.084	0.081	0.085	0.153	0.117	0.110
Ecological regression	0.164	0.113	0.102	0.288	0.293	0.291

Note. The bias for W_j is calculated as $\sum_{i=1}^{n}(\widehat{W}_{ij} - W_{ij})/n$ for $j = 1, 2$, where $\widehat{W}_{ij}$ denotes the in-sample predictions of W_{ij} and W_{ij} is the true value. Similarly, the RMSE is defined as $\sqrt{\sum_{i=1}^{n}(\widehat{W}_{ij} - W_{ij})^2/n}$.

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

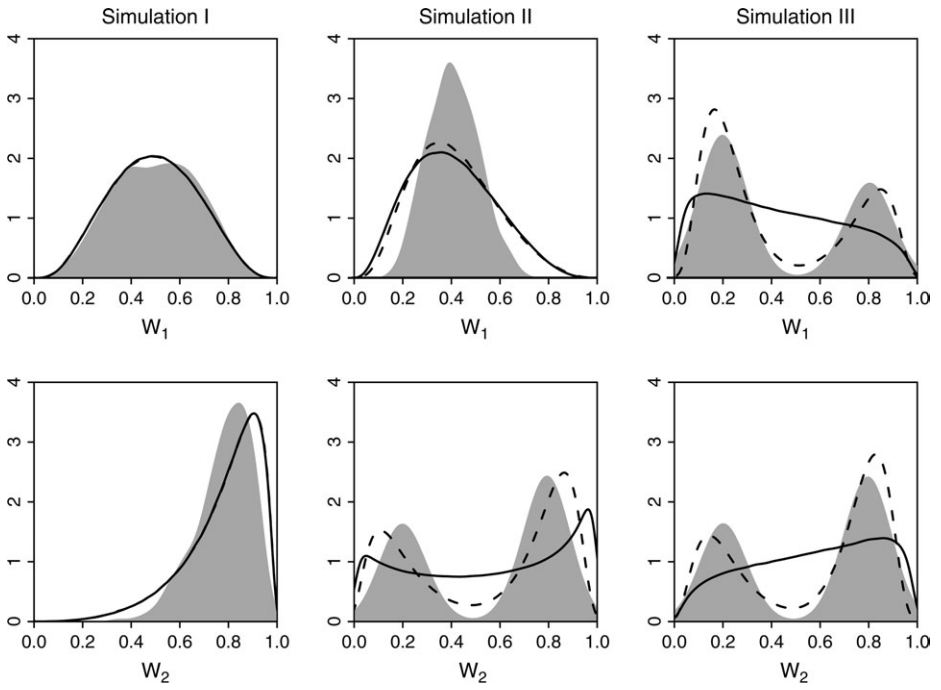


Fig. 2 Out-of-sample predictive performance with different distributions of (W_1, W_2). The true marginal distributions are shown as shaded areas. The solid line represents the estimated density from the parametric model, whereas the dashed line represents that from the nonparametric model.

to the parametric model. This result is not surprising given that this data set is generated using the parametric model. The other two simulations, however, demonstrate the clear advantage of the nonparametric model. The nonparametric model captures the bimodality feature of the marginal distributions, whereas the parametric model fails to approximate the true distribution as expected.

5.1.2 Voter registration in U.S. Southern states

Next, we analyze voter registration data from 275 counties of four Southern states in the United States: Florida, Louisiana, North Carolina, and South Carolina. This data set is first studied by King (1997) and subsequently analyzed by others (e.g., King, Rosen, and Tanner 1999; Wakefield 2004b). For each county, X_i represents the proportion of black voters, Y_i denotes the registration rate, and W_{i1} and W_{i2} represent the registration rates of black and white voters. In this example, the true values of W_{i1} and W_{i2} are known, which allows us to compare the performance of our method with that of existing models.

Figure 3 presents a graphical summary of the data. The upper-left panel plots the true values of W_{i1} and W_{i2}. The registration rates among white voters are high in many counties, with an average of 86%. In contrast, black registration rates are much lower, with an average of 56%. The sample variances of registration rates are 0.044 and 0.024 for black and white voters, respectively. The other two graphs in the upper panel are the scatterplots of the registration rates and the proportions for black and white voters. In this data set, the correlation between X and W_1 is -0.08, whereas the correlation between X and W_2 is only 0.01, implying minor contextual effects.

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

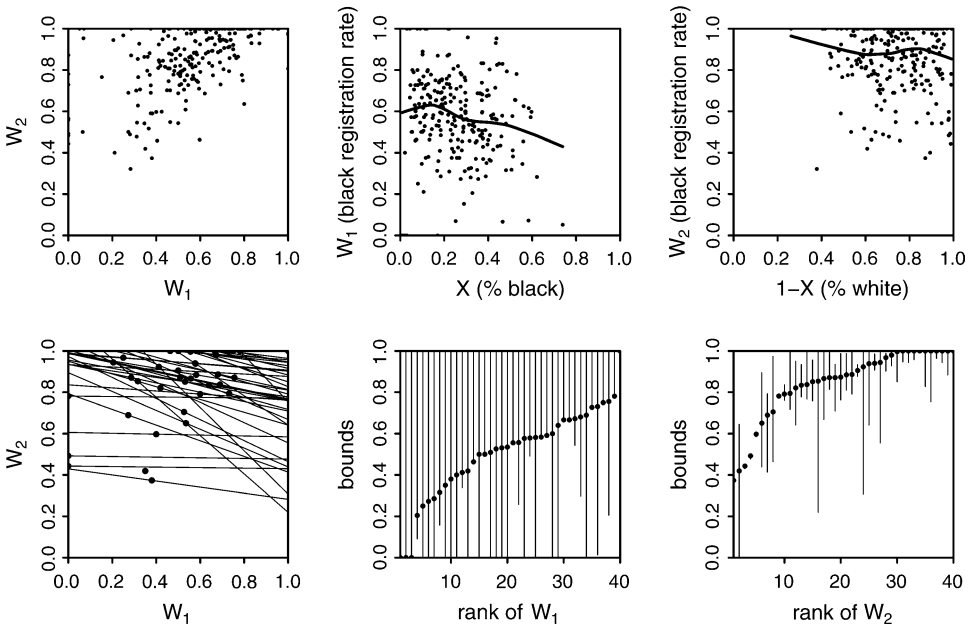


Fig. 3 Summary of the voter registration data from four U.S. Southern states. The upper-left graph is the scatterplot of the true values of W_{i1} and W_{i2}. The upper-middle graph is the scatterplot of black registration rate, W_{i1}, and the ratio of black voters, X_i. The solid line represents a LOWESS curve. The upper-right graph presents the same figure for white voters. The lower-left graph is the tomography plot with the true values indicated as dots. The lower middle and right graphs plot the bounds of W_1 and W_2, respectively.

The lower panel of Fig. 3 presents the tomography plots for a random subset of the counties. The bounds reveal asymmetric information about W_1 and W_2, and they are more informative for W_2 than for W_1. Moreover, for 30% of W_2, the true values are equal to 1. As a result, the true values of the corresponding W_1 lie at the lower end of the bounds. This may pose some difficulty for in-sample predictions, especially for the counties whose bounds are wide.

By treating W_1 and W_2 as unknown, we fit both our parametric and nonparametric models to a subset of 250 counties. We also examine the model performance by adding the individual-level data of the remaining 25 counties. Finally, we compare the results with other methods in the literature, including the ecological regression, the linear and nonlinear neighborhood models (Freedman et al. 1991), the midpoints of bounds, King's EI model, and Wakefield's hierarchical model. To fit King's EI model, we use the publicly available software, EzI (version 2.7) by Benoit and King, with its default specifications. To fit Wakefield's binomial convolution model, we use his WinBUGS code (Wakefield 2004b), which fits the model based on normal approximation. We specify prior distributions such that the implied prior predictive distribution of W_i is approximately uniform. Specifically, we use $\mu_0 \sim \text{logistic}(0, 1)$, $\mu_1 \sim \text{logistic}(0, 1)$, $\sigma_0^{-2} \sim \mathcal{G}(1, 100)$, and $\sigma_1^{-2} \sim \mathcal{G}(1, 100)$. After 50,000 iterations, we discard the initial 20,000 draws and take every 10th draw.

Table 3 summarizes the in-sample predictive performance. For this data set, our nonparametric model significantly outperforms our parametric model in all three discrepancy measures (bias, RMSE, and mean absolute error) by a magnitude that is much greater than what we have seen in our simulation examples. With the addition of the individual-level

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

Table 3 In-sample predictive performance of various models on voter registration data

	Bias		*RMSE*		*MAE*	
	W_1	W_2	W_1	W_2	W_1	W_2
Without survey data						
Parametric model	−0.080	0.030	0.217	0.074	0.170	0.052
Nonparametric model	0.010	−0.003	0.162	0.048	0.111	0.032
With survey data						
Parametric model	0.035	−0.009	0.176	0.056	0.130	0.038
Nonparametric model	0.038	−0.014	0.149	0.055	0.099	0.030
Other methods						
Ecological regression	−0.059	0.016	0.226	0.156	0.177	0.121
King's EI model	0.093	−0.031	0.175	0.065	0.127	0.041
Wakefield's hierarchical model	0.045	−0.013	0.193	0.064	0.145	0.045
Neighborhood method	0.220	−0.077	0.311	0.182	0.247	0.158
Nonlinear neighborhood method	0.220	−0.077	0.269	0.111	0.224	0.078
Midpoints of bounds	0.099	−0.049	0.185	0.092	0.148	0.057

Note. The bias for W_j is calculated as $\sum_{i=1}^{n}(\widehat{W}_{ij} - W_{ij})/n$ for $j = 1, 2$, where $\widehat{W}_{ij}$ denotes the in-sample predictions of W_{ij}, and W_{ij} is the true value. Similarly, the RMSE is defined as $\sqrt{\sum_{i=1}^{n}(\widehat{W}_{ij} - W_{ij})^2/n}$ and the mean absolute error (MAE) is given by $\sum_{i=1}^{n} |\widehat{W}_{ij} - W_{ij}|/n$.

data, however, the in-sample predictions of the parametric model improve substantially. Furthermore, the predictions of the nonparametric model are also more accurate than those of existing methods in terms of all three discrepancy measures. The performance of King's EI model and Wakefield's model is reasonable, but not as good as that of the nonparametric model. Finally, the neighborhood models do not work well in this application, and simply using the midpoint of a bound as an estimate gives better results than some methods.

For our two models, the posterior predictive distribution serves as a basis for population inferences. Figure 4 compares the out-of-sample predictive performance of our models, with and without the addition of individual-level data. In this application, the true distribution of W_1 and W_2 is unknown, so we approximate it by a kernel smoothing technique using the sample values. The nonparametric model estimates the marginal density of W_2 very well, whereas its density estimate for W_1 is slightly off. This is expected because the bounds of W_2 are more informative than those of W_1. In contrast, the estimated marginal densities based on our parametric model are not accurate. With the addition of the individual-level data, the nonparametric model now recovers the density of W_1 and the density estimation of W_2 is further improved. The parametric model still gives a poor estimate even after adding the individual-level data.

5.2 *Contextual Effects*

Next, we investigate the possibility of correcting contextual effects through a simulation study and an empirical investigation of the data set on race and literacy.

5.2.1 A simulation study

To avoid other confounding issues, we simulate a data set under a parametric assumption. We also assume that a covariate Z is an aggregate-level variable, which is expressed in terms of proportion. We start by generating the logit-transformed values of $(W_{i1}, W_{i2}, X_i, Z_i)$,

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

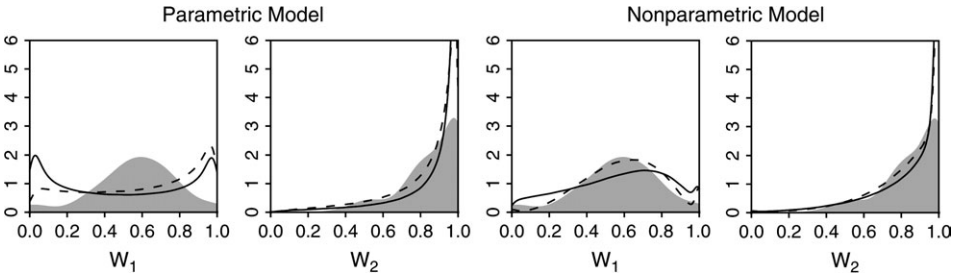


Fig. 4 Out-of-sample predictive performance of selected models on voter registration data. The true density is represented by the shaded area. The solid and dashed lines represent the estimated density without and with the additional survey data information, respectively.

denoted by $(W_{i1}^*, W_{i2}^*, X_i^*, Z_i^*)$ with the sample size of 500. To do this, we first draw Z_i^* independently from a univariate normal distribution with mean -0.85 and variance 0.5. We then compute $W_i^* = BZ_i^* + \varepsilon_{i1}$, where B is a (2×2) matrix with the first diagonal element equal to 0.85, the second diagonal element equal to -0.85, and the off-diagonal elements equal to 0. ε_{i1} is a (2×1) vector independently drawn from a bivariate normal distribution with mean (0, 0), variance (0.5, 0.5), and covariance 0.2. For simplicity, we do not include an intercept.

Next, we construct X^* as a nonlinear function of Z^*. In particular, $X_i^* = 2Z_i^* + 0.5Z_i^{*2} + \varepsilon_{i2}$, where ε_{i2} is a independent draw from a univariate normal distribution with mean 0 and variance 0.5. We then take the inverse-logit transformation of W_i^*, X_i^*, and Z_i^* to obtain W_i, X_i, and Z_i. Finally, applying equation (1), we obtain the value of Y_i. We also generate a spurious covariate $\tilde{Z}$, which is independent of Z, in order to investigate the effect of model misspecification. $\tilde{Z}_i$ is obtained by sampling independently from a normal distribution with mean 0 and variance 0.5 and then taking its inverse-logit transformation.

In this simulation example, X and W are correlated through Z. The sample correlation between X and W_1 is 0.39 and that between X and W_2 is -0.53. Moreover, the average bounds length for W_1 is 0.7 and for W_2 is 0.4, suggesting that W_1 is more coarsened than W_2. Finally, the sample means of W_1 and W_2 are 0.35 and 0.64, respectively.

To examine the performance of various models, we first fit the true model, which is the parametric CCAR model given Z. We then fit four other parametric models (CAR, CCAR given X, CCAR given $\tilde{Z}$, and NCAR). To estimate the proposed Bayesian models, we adopt diffuse prior distributions. In particular, for the parametric CAR model, we use the same prior specifications described in Section 3.1. For the parametric CCAR model, the prior parameters are $B_0 = \mathbf{0}$, $A_0 = I_2$, $\nu_0 = 7$, and $S_0 = 10I_2$; whereas for the parametric NCAR model, our choice of diffuse prior distribution is defined by $\eta_0 = \mathbf{0}$, $\tau_0 = 2$, $\nu_0 = 5$, and $S_0 = 13I_2$.

Table 4 presents the bias and RMSE of the in-sample predictions based for each parametric model. As we expected, when the correct covariate is controlled, the CCAR model yields the smallest bias and RMSE. In contrast, incorrectly conditioning on $\tilde{Z}$ results in poor in-sample predictions. In this particular example, since $\tilde{Z}$ is independent from Z, conditioning on $\tilde{Z}$ does not correct the correlation between X and W. Therefore, the CCAR model behaves like the CAR model. In other cases that are not shown here, when $\tilde{Z}$ is correlated with Z, the in-sample predictions could be even further off from the true values when compared to the CAR model. On the other hand, the parametric NCAR model has a reasonable performance even though it does not incorporate covariate information. Since X is not a linear function of Z at the logit scale, the NCAR model assuming a trivariate normal distribution is not properly specified. Nevertheless, the precision of the in-sample

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

Table 4 In-sample predictive performance of the parametric models on simulated data when X and W are independent given Z

	Bias		*RMSE*	
Model	W_1	W_2	W_1	W_2
CCAR given Z	0.017	0.006	0.127	0.067
CAR	−0.023	0.048	0.163	0.125
CCAR given X	0.066	−0.022	0.167	0.085
CCAR given $\tilde{Z}$	−0.025	0.049	0.163	0.127
NCAR	0.037	−0.008	0.158	0.083

Note. The bias for W_j is calculated as $\sum_{i=1}^{n}(\widehat{W}_{ij} - W_{ij})/n$ for $j = 1, 2$, where $\widehat{W}_{ij}$ denotes the in-sample predictions of W_{ij} and W_{ij} is the true value. Similarly, the RMSE is defined as $\sqrt{\sum_{i=1}^{n}(\widehat{W}_{ij} - W_{ij})^2/n}$.

predictions of W_2 based on this model still improves substantially comparing to those based on the CAR model and the misspecified CCAR models.

5.2.2 Race and literacy

In many studies, the CAR assumption is clearly violated. The straightforward solution in this scenario is to use the CCAR framework; however, this approach may not be feasible for various reasons. For instance, lack of knowledge of the underlying processes would leave the CCAR model vulnerable to misspecification. In other situations, the model may be known but the necessary variables may be unavailable. Therefore, it is of practical importance to consider the analysis under the NCAR assumption.

Here, we reexamine a classical ecological inference problem of black illiteracy rates in 1910 in order to assess the performance of the NCAR models. This study is introduced by Robinson (1950), which is the first article to formally examine the fallacy of ecological inference. Using an empirical example, Robinson (1950) demonstrates that there is not necessarily a correspondence between aggregate- and individual-level correlations. The original study is done based on the state-level data with only 48 observations. To better examine this problem, King (1997) coded the county-level data from the paper records of the 1910 census. In this extended data set, there are 1040 counties. The data set includes the proportion of the residents over 10 years of age who are black X_i, the proportion of those who can read Y_i, the county population size N_i, and the true values of the black literacy rate W_1 and the white literacy rate W_2 with sample mean 68% and 92% for W_1 and W_2, respectively.

Following Robinson (1950), we compare the aggregate correlation between race and literacy with its individual-level counterpart. We first calculate the aggregate correlation as the sample correlation between X_i and Y_i for all the counties. The resulting correlation is -0.733, which is very high. Using the true values of W_i and the number of population in each county, we construct a race $\times$ literacy table, which contains the total number of people in each race by literacy category summing over all 1040 counties. Then we compute the Pearson's correlation coefficient for this 2×2 table, which measures the individual correlation between race and literacy. The resulting individual-level correlation is -0.339, indicating only a mild association between being black and illiterate in 1910 when compared with the aggregate correlation. As demonstrated by Robinson (1950), the large gap between the aggregate and individual correlations implies that we cannot simply use the former to infer the latter.

In this data set, the black literacy rate is negatively correlated with the percentage of black population X (the sample correlation is -0.51), whereas the white literacy rate is

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

Table 5 In-sample predictive performance of various models on literacy data when X and W are correlated

	Bias		*RMSE*	
	W_1	W_2	W_1	W_2
CAR models				
Parametric	−0.064	0.013	0.096	0.031
Nonparametric	−0.060	0.012	0.099	0.030
NCAR models				
Parametric	−0.013	−0.001	0.057	0.027
Nonparametric	0.000	−0.007	0.057	0.029
King's EI models				
No covariate	−0.063	0.013	0.093	0.031
With covariate	−0.057	0.016	0.081	0.029
Wakefields's hierarchical model	−0.055	0.010	0.091	0.030
Ecological regression	−0.072	0.016	0.128	0.059
Neighborhood method	0.141	−0.093	0.168	0.135
Nonlinear neighborhood method	0.141	−0.093	0.151	0.128
Midpoints of bounds	0.022	−0.088	0.140	0.165

Note. The bias for W_j is calculated as $\sum_{i=1}^{n}(\widehat{W}_{ij} - W_{ij})/n$ for $j = 1, 2$, where $\widehat{W}_{ij}$ denotes the in-sample predictions of W_{ij} and W_{ij} is the true value. Similarly, the RMSE is defined as $\sqrt{\sum_{i=1}^{n}(\widehat{W}_{ij} - W_{ij})^2/n}$.

only slightly correlated with X (the sample correlation is 0.17). This suggests that the CAR assumption is likely to be violated. Given the presence of the contextual effect, it is of interest to investigate whether models under the CAR assumption will yield a biased estimate of the individual correlation. We also examine whether the NCAR models can reduce such bias. Moreover, since the parametric assumption about the joint distribution of (W_1^*, W_2^*, X^*) is rather strong, we also study whether the precision of in-sample predictions can be improved by using the nonparametric NCAR model. For the purpose of comparison, we also fit the parametric and nonparametric models under the CAR assumption. To estimate the parametric CAR and NCAR models, we use the same diffuse prior specifications as in Section 3.1. For the nonparametric CAR and NCAR models, the corresponding diffuse prior distributions used in the parametric models are used as the base prior distribution of the Dirichlet processes prior. We also use a diffuse prior distribution for the concentration parameter α, that is, $\Gamma(1, 0.1)$. Table 5 presents the results. As expected, the NCAR models outperform the CAR models and the other models in terms of both bias and RMSE.

Finally, we estimate the individual correlation between race and literacy based on the in-sample predictions of our CAR and NCAR models as well as King's EI models and ecological regression. The results are shown in Table 6. The NCAR models perform best, yielding the estimated individual correlations of −0.341 and −0.359, respectively. In particular, the estimate based on the nonparametric model is very close to the true observed correlation −0.339. In contrast, the estimates based on the other models deviate further from the true value.

5.3 *Aggregation Effects*

Although aggregation effects are inherent to ecological inference problems and cannot be remedied by statistical techniques, the analysis below exhibits the amount of missing information present in the extended literacy data set we analyzed in Section 5.2.

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

Table 6 Estimated individual correlations based on different models

	CAR models		*NCAR models*		*King's EI models*		
True value	*Parametric*	*Non-parametric*	*Parametric*	*Non-parametric*	*With covariate*	*No covariate*	*Ecological regression*
−0.339	−0.414	−0.410	−0.359	−0.341	−0.411	−0.412	−0.401

Note. The correlation is calculated as the Pearson Correlation Coefficient from the 2×2 race/literacy ecological tables with in-sample predictions of W_i from each model. The true value represents the correlation coefficient based on the observed values of W_i.

To keep the example simple, we model the literacy rate within the parametric framework, using the CAR assumption. First, the parameters are estimated and the amount of missing information is quantified for the entire data set, without additional survey data. Next, the data set is supplemented with survey data at amounts ranging from 5% to 15% at 5% intervals—the survey data replace the original data at each record, keeping the overall sample size constant at 1040. The survey data are added to random data points, and the simulation is repeated 20 times at each level of supplemental data. This design results in 60 simulations with survey data, plus one simulation without.

The results of the simulations are presented in Table 7. The literacy rate parameter estimates ($\hat{\mu}_1$and $\hat{\mu}_2$) are logit transformed (e.g., the estimated population black literacy rate in the unsupplemented example is 66%). The amount of missing information is greater

Table 7 Parameter estimates and fraction of missing information for varying levels of supplemental data for the race/literacy data set

		Amount of survey data				
Parameters		*0%*	*5%*	*10%*	*15%*	*Sample estimates*
μ_1	est.	0.654	0.705	0.732	0.742	0.823
	s.e.	(0.032)	(0.028)	(0.025)	(0.024)	(0.016)
	miss.	0.626	0.625	0.610	0.577	
μ_2	est.	2.785	2.676	2.626	2.618	2.649
	s.e.	(0.066)	(0.051)	(0.042)	(0.037)	(0.023)
	miss.	0.567	0.581	0.579	0.553	
σ_1	est.	0.236	0.251	0.260	0.269	0.283
	s.e.	(0.021)	(0.019)	(0.018)	(0.018)	(0.012)
	miss.	0.661	0.642	0.616	0.580	
σ_2	est.	0.916	0.761	0.683	0.660	0.564
	s.e.	(0.114)	(0.073)	(0.056)	(0.047)	(0.024)
	miss.	0.643	0.629	0.608	0.575	
ρ	est.	0.271	0.346	0.398	0.404	0.417
	s.e.	(0.112)	(0.074)	(0.059)	(0.051)	(0.026)
	miss.	0.794	0.780	0.767	0.731	
Hypothesis test	stat.	462.8	546.7	635.4	725.1	2094.7
(H_0: $\mu_1 = \mu_2$)	miss.	0.839	0.780	0.737	0.693	

Note. For each parameter, the point estimate (est.), standard error (s.e.), and fraction of missing information (miss.) are provided. For hypothesis test, the likelihood ratio test statistic (stat.) and the fraction of missing information are presented (miss.). The fraction of missing information is given in percentage. Estimates with supplemental data are averaged over 20 simulations. Sample estimates are the results of applying the parametric model to the individual-level data.

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

for the literacy of blacks than of whites because, on average, blacks make up a lower percent of county populations, resulting in weaker bounds. As survey information is added to the data set, the percent of missing information monotonically decreases. Furthermore, the point estimates of the parameter generally become more accurate with increases in supplemental data. However, even with 15% of the data set containing the actual disaggregated data, more than 50% of information is missing for each parameter estimate and the complete-data estimates (right-most column of Table 7) of two parameters (μ_1 and σ_2) lie outside their 95% confidence intervals.

In addition to parameter estimates, we quantify the amount of missing information in hypothesis testing. For this example, the null hypothesis is that the population white literacy rate and black literacy rate are equal, that is, H_0:$\mu_1 = \mu_2$. This restriction is a more general constraint than the "neighborhood" model, in which the literacy rates are equal within each county ($W_{i1} = W_{i2}$). The (observed) likelihood ratio test statistic (double the numerator of equation 11) is presented in the penultimate row of Table 7. The gap between the likelihoods of the constrained and unconstrained parameter estimates grows (suggesting stronger evidence against the null hypothesis) as more individual-level data are available. Although the fraction of missing information for the hypothesis test begins at the relatively large value of 84%, the decline in this fraction over the amount of supplemental data is steeper than for the parameter estimates.

6 Concluding Remarks

In this article, we show that by formulating an ecological inference problem as an incomplete-data problem, the three key factors that influence ecological inference—aggregation, distributional, and contextual effects—can be formally identified. The proposed framework shows that although distributional and contextual effects can be adjusted by statistical methods, it is the data aggregation that causes the fundamental difficulty of ecological inference and makes the statistical adjustment of the other two factors difficult in practice.

We address each of these three factors. First, to deal with distributional effects, we extend our basic parametric model and propose a Bayesian nonparametric model for ecological inference in 2 × 2 tables. The simulation studies and an empirical example demonstrate that in general the nonparametric model outperforms parametric models by relaxing distributional assumptions. Second, we also demonstrate that contextual effects can be addressed under the proposed parametric and nonparametric models in a relatively straightforward manner. In particular, we show that this task can be accomplished even when extra covariate information is not available. Third, although aggregation effects cannot be statistically adjusted, we demonstrate how to quantify the information loss due to data aggregation in ecological inference. We offer computational methods to quantify the amount of missing information in the context of both parameter estimation and hypothesis testing.

It is important to emphasize that when the aggregation effects are too severe and bounds are too wide, any ecological inference models including our proposed methods are likely to fail. In such situations, the comparison of the predictive distribution of Y from the fitted model against its observed marginal distribution may be able to rule out some of the misspecified models, but the data will not contain enough information to nail down the correct model specification.

Finally, the theoretical framework developed in this article applies more generally to $R \times C$ ecological inference problems where $R \geq 2$ and $C \geq 2$. However, since W_i is of higher dimension in these cases, modeling the three factors simultaneously and detecting

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

possible model misspecification are even more challenging tasks for large ecological tables than for the 2×2 tables considered in this article. Although in theory our nonparametric modeling approach can be extended to larger ecological tables, we believe that such a modeling strategy may not work well in practice due to the lack of information in large ecological tables. Strong parametric assumptions may be necessary when making such inferences.

Appendices: Computational Details

Appendix A: *The EM and ECM algorithms*

In this appendix, we describe the EM and ECM algorithms we developed in order to obtain the ML estimates of the proposed models. The algorithm starts with an arbitrary initial value of parameters, $\theta^{(0)}$, and repeats the expectation step (or E-step) and the maximization step (M-step) until a satisfactory degree of convergence is achieved. The ECM algorithm replaces the M-step with the conditional M-steps where the parameters are divided into smaller subsets and each subset is maximized conditional on the current values of other parameters.

A.1 E-step

At the $(t+1)$th iteration, our E-step for the CAR model requires the integration of the complete-data log likelihood, that is, $l_{\text{com}} = \sum_{i=1}^{n} \log f(W_i \mid \zeta)$, with respect to the missing data, W, over its conditional distribution given the observed data, (Y, X), and the value of parameters from the previous iteration, $\theta^{(t)}$. Thus, we compute,

$$Q(\theta \mid \theta^{(t)}) = -\frac{n}{2}\log[\sigma_1\sigma_2(1-\rho^2)] - \frac{1}{2(1-\rho^2)} \times \left[\frac{S_{11}^{(t)} - 2S_1^{(t)}\mu_1 + \mu_1^2}{\sigma_1} + \frac{S_{22}^{(t)} - 2S_2^{(t)}\mu_2 + \mu_2^2}{\sigma_2} - \frac{2\rho(S_{12}^{(t)} - S_1^{(t)}\mu_2 - S_2^{(t)}\mu_1 + \mu_1\mu_2)}{\sqrt{\sigma_1\sigma_2}}\right],$$

where, for $j, j' = 1, 2$, $S_j^{(t)} = \sum_{i=1}^{n} E(W_{ij}^* \mid X_i, Y_i, \theta^{(t)})$ and $S_{jj'}^{(t)} = \sum_{i=1}^{n} E(W_{ij}^* W_{ij'}^* \mid X_i, Y_i, \theta^{(t)})$ are the expected values of sufficient statistics with respect to W_i^* over its conditional distribution, $p(W_i^* \mid Y_i, X_i, \theta^{(t)})$.

Since $S_j^{(t)}$ and $S_{jj'}^{(t)}$ are not available in a closed form, we use the numerical integration to compute the following integral,

$$E[m(W_i^*) \mid Y_i, X_i, \theta^{(t)}] = \frac{\int_{Y_i = \mathcal{M}(X_i, W_i)} m(W_i^*)\kappa(W_i^* \mid \theta^{(t)})\mathrm{d}W_i^*}{\int_{Y_i = \mathcal{M}(X_i, W_i^*)} \kappa(W_i^* \mid \theta^{(t)})\mathrm{d}W_i^*}, \tag{A1}$$

where $m(W_i^*)$ is a function determined by each of the sufficient statistics and $\kappa(W_i^* \mid \theta^{(t)})$ is the kernel of the bivariate normal density function. Equation (A1) can be viewed as a line integral over a scalar field (e.g., Larson, Hostetler, and Edwards 2002). We express W_i^* as a function of a new variable $t \in (0, 1)$, that is, $W_{ij}^*(t) = \text{logit}[tW_{ij}^{\text{U}} + (1-t)W_{ij}^{\text{L}}]$, for $j = 1, 2$, where $W_{ij}^{\text{U}} = \sup W_{ij}$ and $W_{ij}^{\text{L}} = \inf W_{ij}$ are the upper and lower bounds of W_{ij} given in equation (2). Then, we reexpress the integral as,

$$\int_{Y_i = \mathcal{M}(X_i, W_i)} g(W_i^*)\kappa(W_i^* \mid \theta^{(t)})\,\mathrm{d}W_i^* = \int_0^1 g(W_i^*(t))\kappa(W_i^*(t) \mid \theta^{(t)})\left|\left|\frac{\mathrm{d}}{\mathrm{d}t}W_i^*(t)\right|\right|\mathrm{d}t,$$

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

where the integral is taken with respect to $t \in (0, 1)$. This numerical integration can be accomplished using a standard one-dimensional finite numerical integration routine. Furthermore, the accuracy of this numerical integration can be checked by computing $E(W_{i1}^* \mid X_i, Y_i, \theta^{(t)})$ and $E(W_{i2}^* \mid X_i, Y_i, \theta^{(t)})$, separately and then investigating whether equation (1) holds with these conditional expectations.

The E-step of the NCAR model is similar to that of the CAR model. The difference is that the conditional distribution of the missing data given the observed data and the values of the parameters from the previous iteration $p(W_i^* \mid Y_i, X_i, \theta^{(t)})$ are different. In particular, $\kappa(W_i^* \mid \theta^{(t)})$ in equation (A1) is replaced with $\kappa(W_i^* \mid \theta^{(t)}, X_i^*)$, which is the kernel of the bivariate normal distribution with the marginal means equal to $\mu_1^{(t)} + \rho_{13}^{(t)}\sqrt{\sigma_1^{(t)}/\sigma_3^{(t)}}(X_i^* - \mu_3^{(t)})$ and $\mu_2^{(t)} + \rho_{23}^{(t)}\sqrt{\sigma_2^{(t)}/\sigma_3^{(t)}}(X_i^* - \mu_3^{(t)})$, the marginal variances equal to $\sigma_1^{(t)}(1 - \rho_{13}^{(t)^2})$ and $\sigma_2^{(t)}(1 - \rho_{23}^{(t)^2})$, and the correlation coefficient equal to $(\rho_{12}^{(t)} - \rho_{13}^{(t)}\rho_{23}^{(t)})/\sqrt{(1 - \rho_{13}^{(t)^2})(1 - \rho_{23}^{(t)^2})}$.

A.2 M-step

Once the expected values of sufficient statistics are computed, the M-step is a straightforward application of the standard result available in the literature. Namely, for the CAR model, we have

$$\mu_j^{(t+1)} = \frac{S_j^{(t)}}{n}, \quad \sigma_j^{(t+1)} = \frac{T_{jj}^{(t)}}{n}, \quad \rho^{(t+1)} = \frac{T_{12}^{(t)}}{\sqrt{T_{11}^{(t)} T_{22}^{(t)}}}, \tag{A2}$$

where $T_{jj'}^{(t)} = S_{jj'}^{(t)} - S_j^{(t)} S_{j'}^{(t)}/n$ for $j, j' = 1, 2$.

The M-step of the NCAR model is also similar to that of the CAR model. First the two parameters μ_3 and σ_3 do not need to be updated in each iteration because their ML estimates are available in the closed form, that is, $\hat{\mu}_3 = \sum_{i=1}^n X_i^*/n$ and $\hat{\sigma}_3 = \sum_{i=1}^n (X_i^* - \hat{\mu}_3)^2/n$, respectively. Furthermore, μ_1, μ_2, σ_1, σ_2, and ρ_{12} can be updated in the same way as specified in equation (A2). The remaining correlation parameters are updated as $\rho_{j3}^{(t+1)} = (S_{j3}^{(t)} - \hat{\mu}_3 S_j^{(t)})/\sqrt{\hat{\sigma}_3(nS_{jj}^{(t)} - S_j^{(t)^2})}$, where $S_{j3}^{(t)} = \sum_{i=1}^n X_i^* E[W_{ij}^* \mid Y_i, X_i, \theta^{(t)}]$, for $j = 1, 2$. Like the CAR model, the convergence of the NCAR model is monitored in terms of the transformed parameters, μ_j, log σ_j, and 0.5 log $[(1 + \rho_{jj'})/(1 - \rho_{jj'})]$ for all j, j' with $j \neq j'$.

A.3 CM-step

To conduct the CM-steps at the $(t + 1)$th iteration, we first maximize the regression coefficients, β, given the conditional variance $\Sigma^{(t)}$, via

$$\beta^{(t+1)} = \left\{ \sum_{i=1}^n Z_i^{\mathrm{T}} \Sigma^{(t)-1} Z_i \right\}^{-1} \sum_{i=1}^n Z_i^{\mathrm{T}} \Sigma^{(t)-1} E(W_i^* \mid Y_i, Z_i, \theta^{(t)}), \tag{A3}$$

Given $\beta^{(t+1)}$, we update Σ as follows,

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

$$\Sigma^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} E[(W_i^* - Z_i^{\mathrm{T}}\beta^{(t+1)})(W_i^* - Z_i^{\mathrm{T}}\beta^{(t+1)})^{\mathrm{T}} \mid Y_i, Z_i, \theta^{(t)}]. \tag{A4}$$

Finally, when monitoring the convergence, we transform the variance parameters and the correlation parameter so that they are not bounded; we use the logarithm of the variances, that is, $\log \sigma_j$ for $j = 1, 2$, and the Fisher's Z transformation of the correlation parameter, that is, $0.5 \log [(1 + \rho)/(1 - \rho)]$, to improve the normal approximation.

Appendix B: *The MCMC Algorithms*

In this section, we describe our MCMC algorithms to fit the proposed Bayesian parametric and nonparametric models. We focus on the CAR models but similar algorithms can be applied to the NCAR models.

B.1 The parametric model

To sample from the joint posterior distribution $p(W_i^*, \mu, \Sigma \mid Y, X)$, we construct a Gibbs sampler. First, we draw W_i from its conditional posterior density, which is proportional to,

$$\frac{\mathbf{1}\{W_i : Y_i = W_{i1}X_i + W_{i2}(1 - X_i)\}}{\sqrt{2\pi \mid \Sigma \mid} W_{i1}W_{i2}(1 - W_{i1})(1 - W_{i2})} \exp\left[-\frac{1}{2}\{\mathrm{logit}(W_i) - \mu\}^{\mathrm{T}}\Sigma^{-1}\{\mathrm{logit}(W_i) - \mu\}\right], \tag{B1}$$

if $(W_{i1}, W_{i2}) \in (0, 1)$, otherwise the density is equal to 0. Although equation (B1) is not the density of a standard distribution, it has a bounded support because (W_{i1}, W_{i2}) lies on a bounded line segment. Therefore, we can use the inverse-cumulative distribution function method by evaluating equation (B1) on a grid of equidistant points on a tomography line. Given a sample of W_i, we then obtain W_i^* via the logit transformation. Alternatively, Metropolis-Hastings or importance sampling algorithms can be used, although they require separate tuning parameters or target densities for each observation.

Next, we draw (μ, Σ) from their conditional posterior distributions. Note that the observed data (Y_i, X_i) are redundant given W_i^*. The augmented-data conditional posterior distribution has the form of a standard bivariate normal/inverse-Wishart model, $p(\mu, \Sigma \mid W_i^*) \propto p(\mu \mid \Sigma)p(\Sigma)\prod_{i=1}^{n} p(W_i^* \mid \mu, \Sigma)$. This implies that conditioning on W_i^*, sampling (μ, Σ) can be done using the following standard distributions, $\mu \mid W^*, \Sigma \sim \mathcal{N}\left(\frac{\tau_0^2\mu_0 + n\overline{W}^*}{\tau_0^2 + n}, \frac{\Sigma}{\tau_0^2 + n}\right)$, and $\Sigma \mid W^* \sim \mathrm{InvWish}(\nu_0 + n, S_n^{-1})$, where $W^* = W_i^*, \ldots, W_n^*$, $\overline{W}^* = \sum_{i=1}^{n} W_i^*/n$, and $S_n = S_0 + \sum_{i=1}^{n}(W_i^* - \overline{W}^*)(W_i^* - \overline{W}^*)^{\mathrm{T}} + \frac{\tau_0^2 n}{\tau_0^2 + n}(\overline{W}^* - \mu_0)(\overline{W}^* - \mu_0)^{\mathrm{T}}$.

B.2 The nonparametric model

We construct a Gibbs sampler in order to sample from the joint posterior distribution $p(W^*, \mu, \Sigma, \alpha \mid Y)$. First, we independently sample W_i for each i and transform it to obtain W_i^* in the same way as above, but we replace (μ, Σ) with (μ_i, Σ_i) in equation (B1). Then, given the draw of W_i^*, the augmented-data model can be estimated through a multivariate

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

generalization of the density estimation method of Escobar and West (1995). In our Gibbs sampler, we sample (μ_i, Σ_i) given $(\mu^{(i)}, \Sigma^{(i)}, W^*, \alpha)$ for each i and then update α based on the new values of (μ_i, Σ_i).

An application of the usual calculation due to Antoniak (1974) shows that the conditional posterior distribution of (μ_i, Σ_i) given W_i^* is given by the following mixture of Dirichlet processes,

$$(\mu_i, \Sigma_i) \mid \mu^{(i)}, \Sigma^{(i)}, W_i^* \sim q_0 G_i(\mu_i, \Sigma_i) + \sum_{j=1, j\neq i}^{n} q_j \, \delta_{(\mu_j, \Sigma_j)}(\mu_i, \Sigma_i),$$

where $G_i(\mu_i, \Sigma_i)$ is the posterior distribution under G_0 which is a normal/inverse-Wishart distribution with components,

$$\mu_i \mid \Sigma_i \sim \mathcal{N}\left(\frac{\tau_0^2\mu_0 + W_i^*}{\tau_0^2 + 1}, \frac{\Sigma_i}{\tau_0^2 + 1}\right),$$

$$\Sigma_i \sim \text{InvWish}\left[\nu_0 + 1, \left\{S_0 + \frac{\tau_0^2}{\tau_0^2 + 1}(W_i^* - \mu_0)(W_i^* - \mu_0)^{\mathrm{T}}\right\}^{-1}\right].$$

Next, following West, Müller, and Escobar (1994), we derive the weights q_0 and q_j by computing the marginal (augmented data) likelihood $p(W_i^* \mid \mu_i, \Sigma_i)$ and $p(W_i^* \mid \mu_j, \Sigma_j)$, respectively,

$$q_0 \propto \alpha \, \frac{\tau_0^2 \Gamma\left(\frac{\nu_0+1}{2}\right)}{\pi(\tau_0^2 + 1)\Gamma\left(\frac{\nu_0-1}{2}\right)} \mid S_0 \mid^{-1/2} \left\{1 + \frac{\tau_0^2}{\tau_0^2 + 1}(W_i^* - \mu_0)^{\mathrm{T}} S_0^{-1}(W_i^* - \mu_0)\right\}^{-(\nu_0+1)/2},$$

$$q_j \propto \mid \Sigma_j \mid^{-1/2} \exp\left\{\frac{1}{2}(W_i^* - \mu_j)^{\mathrm{T}} \Sigma_j^{-1}(W_i^* - \mu_j)\right\} \quad \text{for } j = 1, \ldots, n, \text{and } j \neq i,$$

where $\sum_{j=0, j\neq i}^{n} q_j = 1$. q_0 is proportional to the bivariate t density with $(\nu_0 - 1)$ degrees of freedom, the location parameter μ_0, and the scale matrix $S_0(1 + \tau_0^2)/\{\tau_0^2(\nu_0 - 1)\}$. q_j is proportional to the bivariate normal density with mean μ_j and variance Σ_j.

Given these weights, we can approximate $p(\mu, \Sigma \mid W^*)$ via a Gibbs sampler by sampling (μ_i, Σ_i) given $(\mu^{(i)}, \Sigma^{(i)}, W_i^*)$ for each i. This step creates clusters of units where some units share the same values of the population parameters. At a particular iteration, we have $J \leq n$ clusters each of which has n_j units with $\sum_{j=1}^{J} n_j = n$. Note that the number of clusters J can vary from one iteration to another. Bush and MacEachern (1996) recommend adding the "remixing" step to prevent the Gibbs sampler from repeatedly sampling a small set of values. In our application, we update the new values of the parameters (μ_i, Σ_i) by using the newly configured cluster structure. That is, for each cluster j, we update the parameters with $(\tilde{\mu}_j, \tilde{\Sigma}_j)$ by drawing them from the following conditional distribution,

$$\tilde{\mu}_j \mid \tilde{\Sigma}_j, \{W_i^* : i \in j\text{th cluster}\} \sim \mathcal{N}\left(\frac{\tau_0^2\mu_0 + n_j\overline{W}_j^*}{\tau_0^2 + n_j}, \frac{\tilde{\Sigma}_j}{\tau_0^2 + n_j}\right),$$

$$\tilde{\Sigma}_j \mid \{W_i^* : i \in j\text{th cluster}\} \sim \text{InvWish}(\nu_0 + n_j, S_{n_j}^{-1}),$$

where $S_{n_j} = S_0 + \sum_{i\in j\text{th cluster}}^{n_j}(W_i^* - \overline{W}_j^*)(W_i^* - \overline{W}_j^*)^{\mathrm{T}} + \frac{\tau_0^2 n_j}{\tau_0^2 + n_j}(\overline{W}_j^* - \mu_0)(\overline{W}_j^* - \mu_0)^{\mathrm{T}}$ and $\overline{W}_j^* = \sum_{i\in j\text{th cluster}}^{n_j} W_i^*/n_j$. Given these new draws, we set $\mu_i = \mu_j^*$ and $\Sigma_i = \Sigma_j^*$ for each i that belongs to the jth cluster.

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

Finally, to update α, we use the algorithm developed by Escobar and West (1995). Namely, the conditional posterior distribution of α has the form of the following gamma mixture,

$$\alpha \mid \eta, J \sim \omega\mathcal{G}(a_0 + J, b_0 - \log\eta) + (1 - \omega)\mathcal{G}(a_0 + J - 1, b_0 - \log\eta),$$

where $\omega = (a_0 + J - 1)/\{n(b_0 - \log \eta)\}$, and η is a latent variable that follows a beta distribution, $\mathcal{B}(\alpha + 1, J)$. This completes one cycle of our Gibbs sampler.

Funding

National Science Foundation (SES–0550873); Princeton University Committee on Research in the Humanities and Social Sciences.

References

Achen, C. H., and W. P. Shively. 1995. *Cross-level inference.* Chicago, IL: University of Chicago Press.

Antoniak, C. E. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* 2:1152–74.

Benoit, Kenneth and Gary King. 2003. EzI: A(n easy) program for ecological inference. Cambridge, Mass.: Harvard University. Available from: http://gking.harvard.edu. (accessed August 8, 2007).

Brown, P. J., and C. D. Payne. 1986. Aggregate data, ecological regression, and voting transitions. *Journal of the American Statistical Association* 81:452–60.

Burden, B. C., and D. C. Kimball. 1998. A new approach to the study of ticket splitting. *American Political Science Review* 92:533–44.

Bush, C. A., and S. N. MacEachern. 1996. A semiparametric Bayesian model for randomized block designs. *Biometrika* 83:275–85.

Cho, W. K. T. 1998. Iff the assumption fits. . .: A comment on the King ecological inference solution. *Political Analysis* 7:143–63.

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* 48:152–71.

Copas, J., and S. Eguchi. 2005. Local model uncertainty and incomplete-data bias. *Journal of the Royal Statistical Society, Series B (Methodological)* 67:459–513.

Cross, P. J., and C. F. Manski. 2002. Regressions, short and long. *Econometrica* 70:357–68.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* 39:1–37.

Dey, D., P. Müller, and D. Sinha, eds. 1998. *Practical nonparametric and semiparametric Bayesian statistics.* New York: Springer-Verlag Inc.

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

Escobar, M. D., and M. West. 1995. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 90:577–88.

Ferguson, T. S. 1973. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1:209–30.

Freedman, D. A., S. P. Klein, J. Sacks, C. A. Smyth, and C. G. Everett. 1991. Ecological regression and voting rights (with discussion). *Evaluation Review* 15:673–816.

Freedman D. A., M. Ostland, M. R. Roberts, and S. P. Klein. 1998. Review of "A Solution to the Ecological Inference Problem." *Journal of the American Statistical Association* 93:1518–22.

Gelman, A., D. K. Park, S. Ansolabehere, P. N. Price, and L. C. Minnite. 2001. Models, assumptions and model checking in ecological regressions. *Journal of the Royal Statistical Society, Series A* 164:101–18.

Gill, J., and G. Casella. 2006. Markov chain Monte Carlo methods for models with nonparametric priors. Technical report, University of California, Davis.

Goodman, L. 1953. Ecological regressions and behavior of individuals. *American Sociological Review* 18: 663–6.

Grofman, B. 1991. Statistics without substance: A critique of Freedman et al. and Clark and Morrison. *Evaluation Review* 15:746–69.

Heitjan, D. F., and D. B. Rubin. 1991. Ignorability and coarse data. *The Annals of Statistics* 19:2244–53.

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press

Herron, M. C., and K. W. Shotts. 2004. Logical inconsistency in EI-based second stage regressions. *American Journal of Political Science* 48:172–83.

Imai, K., and G. King. 2004. Did illegal overseas absentee ballots decide the 2000 U.S. presidential election? *Perspectives on Politics* 2:537–49.

Imai, K., Y. Lu, and A. Strauss. eco: R package for ecological inference in 2 × 2 tables. *Journal of Statistical Software* (forthcoming).

Judge, G. G., D. J. Miller, and W. K. T. Cho. 2004. An information theoretic approach to ecological estimation and inference. In *Ecological inference: New methodological strategies*, ed. G. King, O. Rosen, and M. Tanner, 162–87. Cambridge: Cambridge University Press.

King, G. 1997. *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data.* Princeton, NJ: Princeton University Press.

———. 1999. Comment on "review of 'a solution to the ecological inference problem'." *Journal of the American Statistical Association* 94:352–5.

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

King, G., O. Rosen, and M. A. Tanner, eds. 2004. *Ecological inference: New methodological strategies.* Cambridge: Cambridge University Press.

Kong, A., X.-L. Meng, and D. L. Nicolae. 2005. Quantifying relative incomplete information for hypothesis testing in statistical and genetic studies. Unpublished manuscript, Department of Statistics, Harvard University.

Larson, R., R. P. Hostetler, and B. H. Edwards. 2002. *Calculus: Early transcendental functions.* 3rd ed. Boston, MA: Houghton Mifflin Company.

Meng, X.-L., and D. B. Rubin. 1991. Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association* 86:899–909.

———. 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* 80:267–78.

Mukhopadhyay, S., and A. E. Gelfand. 1997. Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association* 92:633–9.

Neyman, J., and E. L. Scott. 1948. Consistent estimation from partially consistent observations. *Econometrica* 16:1–32.

Orchard, T., and M. A. Woodbury 1972. A missing information principle: Theory and applications. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability* 1:697–715.

Robinson, W. S. 1950. Ecological correlations and the behavior of individuals. *American Sociological Review* 15:351–7.

Rosen, O., W. Jiang, G. King, and M. A. Tanner. 2001. Bayesian and frequentist inference for ecological inference: The R × C case. *Statistica Neerlandica* 55:134–56.

van Dyk, D. A., X.-L. Meng, and D. B. Rubin. 1995. Maximum likelihood estimation via the ECM algorithm: Computing the asymptotic variance. *Statistica Sinica* 5:55–75.

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

———. 2004b. Prior and likelihood choices in the analysis of ecological data. In *Ecological inference: New methodological strategies*, ed. Gary King, Ori Rosen, and Martin Tanner, 13–50. Cambridge: Cambridge University Press.

West, M., P. Müller, and M. D. Escobar. 1994. Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of uncertainty: A tribute to D. V. Lindley*, ed. A. F. M. Smith and P. R. Freedman, 363–86. London: John Wiley & Sons.

https://doi.org/10.1093/pan/mpm017 Published online by Cambridge University Press