

An Incomplete Data Approach to the Ecological Inference Problem

Kosuke Imai*

Department of Politics, Princeton University

Ying Lu†

The Institute for Quantitative Social Science, Harvard University

First Draft: August 7, 2005

This Draft: August 7, 2005

Abstract

In this paper, we propose to formulate ecological inference as a coarse data problem where only a subset of the complete-data sample space is observed. Applying the related assumptions and theoretical results of Heitjan and Rubin (1991), we formally identify three key factors that affect ecological inference; distributional, contextual and aggregation effects. Different modeling strategies are discussed to deal with distributional and contextual effects. While aggregation effects cannot be statistically adjusted, we show how to formally quantify the magnitude of such effects through the use of the Expectation-Maximization algorithm. The paper concludes with simulations and empirical applications that assess the performance of the proposed models. C-code used to implement the proposed method is available with easy-to-use R interface.

Key Words: Coarse data, Contextual effects, Data augmentation, *EM* algorithm, Missing information principle, Nonparametric Bayesian Modeling.

*Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609-258-6610, Fax: 973-556-1929, Email: kimai@Princeton.Edu, URL: <http://imai.princeton.edu>

†Postdoctoral Fellow, The Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138. Phone: 617-496-2031, Fax: 617-496-5149, Email: ylu@Iq.Harvard.Edu

1 Introduction

Ecological inference refers to the “inferences about individual behavior drawn from data about aggregates” (Freedman, 1999, p.4027). Such cross-level inferences are frequently conducted in the social sciences and epidemiology when only aggregate data are available to researchers. For example, the analysis of voting behavior by different racial groups bears important implications for Congressional redistricting in the United States (e.g., Freedman *et al.*, 1991; Grofman, 1991; Lichtman, 1991). In such an analysis, due to the confidentiality of the ballots, questions about racial voting behavior are often answered by examining, at the level of electoral districts, the correlation between voting data from the official election results and racial composition data from the Census data (e.g., Shaw, 1997; Hajnal *et al.*, 2002). Political scientists also apply ecological inference to investigate the behavioral mechanism of “split ticketing” where voters cast ballots for candidates from different parties in Congressional and Presidential elections (e.g., Alesina and Rosenthal, 1995; Burden and Kimball, 1998; Fiorina, 1996).

In sociology, due to the sensitivity of subjects, researchers sometimes lack access to individual level information. A well-known example is the research that investigates whether suicide is encouraged by the social conditions surrounding protestants (e.g., Durkheim, 1897; Neeleman and Lewis, 1999). In such studies, researchers examine the regional correlation between suicide rates and the size of Protestant population. In other cases, the limitation of the historical records makes the analysis of individual level data impossible. For example, many have studied the question of who gave support for the Nazi Party in pre-war Germany by analyzing the aggregate electoral data (e.g., King *et al.*, 2004a; Lohmoller *et al.*, 1985; O’Loughlin, 2000).

In epidemiology, ecological inference is also relevant when assessing the disease risk factors by examining aggregate relationships between the risk exposure rate and disease clusters. The exam-

ples include ecological studies about the effects of water contamination on cholera transmission (e.g., Snow, 1855) and about the effects of the residential radon exposure on lung cancer incidences (e.g., Goldsmith, 1999; Darby *et al.*, 2001).

While its applications are abundant in a variety of scientific disciplines, the difficulty of ecological inference is that the observed correlation at the aggregate level does not necessarily agree with the individual-level relationship. Using an example of literacy rates across different racial groups, Robinson (1950) powerfully illustrated this “ecological fallacy.” Since then, a number of statisticians and methodologists have proposed various methods for the ecological inference problem. Duncan and Davis (1953) showed how to derive the bounds on unknown quantities of interest from aggregate data. Goodman (1953, 1959) developed the regression-based approach to ecological inference, which had gained popularity among applied researchers in the next several decades (e.g., Achen and Shively, 1995; Freedman *et al.*, 1991). Recent years have witnessed a growing number of new methods based on modern statistical techniques including Markov chain Monte Carlo (e.g., King, 1997; King *et al.*, 1999; Johnston and Pattie, 2000; Rosen *et al.*, 2001; Wakefield, 2004; King *et al.*, 2004b; Imai and Lu, 2004). At the same time, the appropriateness of the particular assumptions underlying these models is often debated (e.g., Freedman *et al.*, 1998; Cho, 1998; King, 1999).

In this paper, we contribute to this fast growing literature by developing a new approach to the ecological inference problem that is based on an incomplete-data perspective. In particular, we show that ecological inference can be viewed as a special case of the coarse data problem where only a subset of the complete-data sample space is observed. We then discuss how the general assumptions and theoretical results of Heitjan and Rubin (1991) can be applied to the ecological inference problem. The main advantage of the proposed approach is that it formally

defines the key assumptions and identifies challenges for ecological inference within the general framework of inference with missing data (Rubin, 1976; Little and Rubin, 1987). In particular, we formally identify three key factors that affect ecological inference; distributional, contextual and aggregation effects.

Within this formal framework of coarse data, the method of data augmentation serves as a natural and general modeling strategy for ecological inference (Tanner and Wong, 1987). Different modeling strategies are discussed to deal with distributional and contextual effects. While aggregation effects cannot be statistically adjusted, we show how to formally quantify the magnitude of such effects through the use of the Expectation-Maximization (*EM*) algorithm (Dempster *et al.*, 1977). Using both simulated and real data sets, we illustrate how to quantify the amount of information loss due to aggregation effects, to examine the impact of model misspecification due to distributional effects, and to cope with potential contextual effects in ecological inference. C-code used to implement the proposed method is available with easy-to-use R interface (Imai and Lu, 2005).

The rest of this paper proceeds as follows. In Section 2, we propose a theoretical framework for ecological inference from an incomplete data perspective, and present the conditions under which valid inferences can be made. In Section 3, we illustrate the advantage of using the method of data augmentation as a general modeling strategy for ecological inference. In Section 4, we analyze simulated and real datasets to demonstrate the advantages of our modeling strategy. Finally, Section 5 gives concluding remarks.

2 An Incomplete Data Framework for Ecological Inference

In this section, we introduce an incomplete data approach to the ecological inference problem. 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 the stochastic nature of the coarsening mechanism can be ignored when making ecological inference via either Bayesian or likelihood approach. We demonstrate that this approach clarifies and formally defines the modeling assumptions necessary for ecological inference. The proposed framework also reveals the fundamental difficulty inherent in cross-level inference.

2.1 Notations and Goals of Ecological Inference for 2×2 Tables

First, we introduce the notations of ecological inference that are used throughout this paper. Our focus is on statistical issues that arise in analyzing 2×2 ecological tables where both the outcome variable y and the predictor x are binary indicator variables. In ecological inference, researchers observe only the aggregate summaries of these two variables; i.e., the proportion or count of individuals with $y = 1$ and the proportion or count of individuals with $x = 1$. Given the aggregate data, the relationship between y and x at the individual level is of interest; e.g., the proportion of individuals with $x = 1$ and $y = 1$. Although larger $r \times c$ ecological tables where $r > 2$ and/or $c > 2$ are encountered in applied research, a clear understanding of 2×2 tables serves as a starting point for examining more complex cases.

Table 1 shows our notations of ecological inference for 2×2 tables. Suppose there are n geographic units. We assume that in each unit i , such a 2×2 ecological table is separately obtained. For each unit i , we observe the proportion of individuals with $y = 1$ and denote it by Y_i . We also observe the proportion of individuals with $x = 1$, for which we use the notation X_i .

| | | | |
|---------|--------------|--------------|-----------|
| | $x = 1$ | $x = 0$ | |
| $y = 1$ | W_{1i} | W_{2i} | Y_i |
| $y = 0$ | $1 - W_{1i}$ | $1 - W_{2i}$ | $1 - Y_i$ |
| | X_i | $1 - X_i$ | |

Table 1: Notations of 2×2 Ecological Table in Terms of Proportions. X_i, Y_i, W_{1i} , and W_{2i} are proportions, and hence lie between 0 and 1. The unit of observation is typically a geographical unit and is denoted by i .

The four internal cells are not observed. W_{1i} represents the proportion of individuals with $y = 1$ among all individuals with $x = 1$. Similarly, W_{2i} represents the proportion of individuals with $y = 1$ among all individuals with $x = 0$. While both W_{1i} and W_{2i} are not observed, they follow a key deterministic relationship,

$$Y_i = W_{1i}X_i + W_{2i}(1 - X_i). \tag{1}$$

That is, Y_i is the observed weighted average of the two unknown variables, W_{1i} and W_{2i} , with X_i and $1 - X_i$ being the observed weights.

The goals of ecological inference for 2×2 tables 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 distributions of W_1 and W_2 , or the distributions themselves. Second, since the internal cells are not observed, the estimation of the values of W_{1i} and W_{2i} for each unit i is also important. We call the former *population ecological inference*, while the latter is referred to as *in-sample ecological inference*. In studies of racial voting, for example, in-sample ecological inference rather than population inference is often emphasized. However, in many other studies such as the ones that assess disease risk factors through ecological data, population ecological inference is of primary interest.

We emphasize that for in-sample ecological inference, W_{1i} and W_{2i} should not be treated as unknown parameters. If we consider W_{1i} and W_{2i} as parameters to be estimated, then we must

estimate $2n$ parameters based on n observations, and no informational gain results from obtaining additional observations. Instead, each new observation creates two additional parameters to estimate. This is then a special case of the incidental parameter problem where no consistent estimators can be constructed for W_{1i} and W_{2i} (e.g., Neyman and Scott, 1948). Hence, W_{1i} and W_{2i} must be viewed as realizations to be predicted. In contrast, population parameters can be estimated if we treat the incidental parameters as “nuisance parameters.” Therefore, the distinction between in-sample and population inferences is critical for understanding the statistical properties and evaluating the performance of various ecological inference models.

2.2 Ecological Tables as Coarse Data

Here, we 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. Coarse data can be neither entirely unknown 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 problem in 2×2 tables, the vector of internal cells $W_i = (W_{1i}, W_{2i})$ 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_{1i} and W_{2i}

$$W_{1i} \in \left[\max \left(0, \frac{X_i + Y_i - 1}{X_i} \right), \min \left(1, \frac{Y_i}{X_i} \right) \right], \quad (2)$$

$$W_{2i} \in \left[\max \left(0, \frac{Y_i - X_i}{1 - X_i} \right), \min \left(1, \frac{Y_i}{1 - X_i} \right) \right]. \quad (3)$$

While these intervals reveal the possible values that W_{1i} and W_{2i} could take, they are often too

wide to be informative for the purposes of applied researchers.

From the perspective of incomplete data, Y_i can be viewed as a coarse version of the unobserved data $W_i = (W_{1i}, W_{2i})$ where equation (1) characterizes the relationship between Y_i and W_i . The value of X_i determines how the original data W_i are coarsened. For example, if $X_i = 1$, then W_{1i} is known and equals to Y_i , but W_{2i} is considered as completely missing. On the other hand, if $X_i = 0$, then W_{2i} is known exactly, but W_{1i} is completely missing. Moreover, if X_i takes a value closer to 1, bounds are likely to be narrow for W_{1i} and wide for W_{2i} , implying that W_{2i} is coarsened more than W_{1i} . Therefore, if X_i is viewed as a random variable, equation (1) characterizes the stochastic data coarsening process.

Formally, we characterize the ecological inference problem in 2×2 tables as a coarse data problem in the following manner. The random vector $W = (W_1, W_2)$ represents the variables of interest with a sample space $\Xi = [0, 1] \times [0, 1]$. W is assumed to be distributed with density $f(W \mid \theta)$ and unknown parameter θ . The goal of population ecological inference is to make inferences about θ . Y , the weighted sum of W_1 and W_2 , is viewed as a coarsened version of W . Each observation Y maps onto the subspace of Ξ . Hence, the sample space defined by Y is the power set, 2^{Ξ} . If we view X as a random variable with sample space Γ , the conditional distribution of X given unobserved data W and unknown parameter γ is denoted by $h(X \mid W, \gamma)$. It is the distribution of X that determines the coarsening mechanism.

2.3 Theoretical Framework

In this section, we place ecological inference within the formal theoretical framework of coarse data developed by Heitjan and Rubin (1991). We begin by defining the conditional distribution

of Y given (W, X) , which is a degenerate function,

$$r(Y|W, X, \theta, \gamma) = r(Y|W, X) = \begin{cases} 1 & \text{if } Y = Y(W, X) \\ 0 & \text{if } Y \neq Y(W, X) \end{cases}, \quad (4)$$

where $Y(X, W) = XW_1 + (1 - X)W_2$. Then, the conditional distribution of observed data Y and the coarsening variable X given unobserved data W can be expressed as follows,

$$k(Y, X | W, \gamma) = r(Y | W, X) h(X | W, \gamma). \quad (5)$$

Finally, using the definitions of these conditional distributions, the observed-data likelihood function of coarse data can be written as,

$$L_{obs}(\theta, \gamma | Y, X) = \int_{\Xi} k(Y, X | W, \gamma) f(W | \theta) dW \quad (6)$$

$$= \int_{\Xi} r(Y | W, X) h(X | W, \gamma) f(W | \theta) dW, \quad (7)$$

where $f(W | \theta)$ can be seen as the complete-data likelihood.

To make inferences based on $L_{obs}(\theta | Y, X)$, we must specify the sampling distribution of unobserved data $f(W | \theta)$ as well as the conditional distribution of the weight variable $h(X | W, \gamma)$ which characterizes the coarsening mechanism. In ecological inference, this incomplete-data framework allows us to formally identify the three key factors highlighted by Imai and Lu (2004). The specification of $f(W | \theta)$ relates to *distributional effects* because it models the distribution of missing data, while $h(X | W, \gamma)$ needs to be specified in order to model *contextual effects*. Finally, *aggregation effects* refer to the general loss of information due to data coarsening.

Since W cannot be directly observed, the specification of $h(X | W, \gamma)$ and $f(W | \theta)$ poses a serious challenge and is fundamentally untestable in ecological inference. Thus, it is desirable to consider the general conditions under which strong assumptions can be avoided. First, we state the condition under which the stochastic coarsening mechanism can be ignored; i.e., the

condition under which we need not specify $h(X | W, \gamma)$. In ecological inference, this corresponds to the condition under which contextual effects can be ignored. Heitjan and Rubin (1991) formally defines this condition, and call it *coarsened at random* (CAR) as a general formulation of missing at random (MAR) in the literature on inference with missing data (Rubin, 1974, 1976). In general, because the coarse data are partially observed and never completely missing, there exists no condition that is equivalent to the assumption of missing completely at random (MCAR). In the context of ecological inference, this can be seen from the fact that the bounds contain the information about W .

The formal definition of CAR is given by Heitjan and Rubin (1991),

DEFINITION 1 (COARSENEDED AT RANDOM (CAR)) *The data Y is coarsened at random (CAR) if, for the fixed observed value of Y and for each value of γ , $\int_{\Gamma} k(Y, X | W, \gamma) dX$ takes the same value for all $W \in Y(W, X)$, that is for all values of W that are consistent with the observed data Y .*

Under CAR, if θ and γ are disjoint parameters, the inference about θ does not depend on the nuisance parameter γ and hence the specification of $h(X | W, \gamma)$ can be ignored. Heitjan and Rubin (1991) also shows that CAR is the weakest condition under which it is appropriate to ignore the data coarsening mechanism when making valid inference with coarse data.

A sufficient condition for Y to be CAR is that X and W are independent; i.e., $h(X | W, \gamma) = h(X | \gamma)$ for $W \in Y(W, X)$. When θ and γ are disjoint parameters, $h(X | \gamma)$ does not affect the estimation of θ given the observed likelihood $L_{obs}(\theta | Y, X)$. As a consequence, we can simplify the observed-data likelihood function of equation (7) as,

$$L_{obs}(\theta | Y, X) = \int_{\Xi} r(Y | W, X, \theta) f(W | \theta) dW. \quad (8)$$

Heitjan and Rubin (1991) refers to a related situation as data grouping where the coarsening process is random and the coarse version of W can be expressed as $Y = Y(W)$. For data grouping, the conditional distribution of Y given W is also a degenerate function,

$$r(Y | W, \theta) = r(Y | W) = \begin{cases} 1 & \text{if } Y = Y(W), \\ 0 & \text{if } Y \neq Y(W) \end{cases}. \quad (9)$$

Under CAR, ecological data are similar to grouped data except that the coarse version of (W_1, W_2) , i.e., Y_i , is further determined by the values of X .

On the other hand, when the CAR assumption does not hold, the values of W_{1i} and W_{2i} can affect their coarse representations. In particular, as specified in equations 2 and 3, the upper and lower bounds of W_{1i} (W_{2i}) are both monotonically decreasing (increasing) function of X_i . Hence when X and W_1 (W_2) are not independent, the value of W_{1i} (W_{2i}) tends to affect the location of its bound interval. In this case, if we disregard such coarsening process, the resulting inference is likely to be biased.

In some situations, researchers know a priori that W and X are not independent, and wish to model the coarsening mechanism through controlling some observed covariates Z such that the CAR assumption holds given Z . We refer to this situation as *conditionally coarsened at random* or CCAR. For example, in the context of racial voting, turnout rates are likely to be correlated with racial composition through the income and education levels of electoral districts. Formally, this assumption is that W and X are conditionally independent given Z , i.e., $h(X | W, Z, \gamma) = h(X | Z, \gamma)$. If the assumption holds, the data are CAR after conditioning on Z , and the (conditional) observed-data likelihood $L_{obs}(\theta | Y, X, Z)$ can be written as,

$$L_{obs}(\theta | Y, X, Z) = \int_{\Xi} r(Y | W, X, \theta) f(W | Z, \theta) dW. \quad (10)$$

Next, we consider a scenario where CAR is known to be violated and no covariate Z is available

such that CCAR holds. We refer to this situation as *not coarsened at random* or NCAR. In this case, we must directly model the data coarsening mechanism and specify the joint distribution $g(X, W | \theta, \gamma) = f(W | \theta)h(X | W, \gamma)$. Then, the observed-data likelihood can be written as,

$$L_{obs}(\theta, \gamma | Y, X) = \int_{\Xi} r(Y | W, X) g(X, W | \theta, \gamma) dW. \quad (11)$$

For example, one might assume that W and X are jointly normally distributed. In that case, the conditional distribution $h(X | W, \gamma)$ is also normal.

In sum, starting with the general formulation of observed-data likelihood in equation (7), there are three ways to proceed for the purpose of ecological inference. First, if we assume no contextual effect, i.e., X and W are independent, then the CAR condition holds and the data coarsening mechanism can be ignored. Valid inferences can then be made using the simplified likelihood function in equation (8). Secondly, if we assume that X and W are conditionally independent given some observed covariates Z (i.e., CCAR), then the correct observed data likelihood is given in equation (10) and one must specify the conditional distribution $f(W | Z, \theta)$. Finally, if the CAR assumption cannot be satisfied at all (i.e., NCAR), then we must base our inference on equation (11) and model the joint distribution, $g(X, W | \theta, \gamma)$.

Therefore, an important task for applied researchers is to specify the distribution functions necessary for each of the three modeling strategies outlined above. Since W is not directly observed, the related distributions $f(W | \theta)$, $f(W | Z, \theta)$, and $g(W, X | \theta)$ are essentially untestable. In general, we recommend using a nonparametric or semiparametric specification to avoid relying on strong parametric assumptions. Moreover, the third (NCAR) strategy should often be preferred over the second (CCAR) strategy because the specification of the conditional distribution $f(W | Z, \theta)$ is often much more difficult than that of the joint distribution $g(W, X | \theta, \gamma)$. The former necessarily involves the variable selection issue related to Z as well as the distributional assumption

concerning $f(W | Z, \theta)$. When the dimension of Z is large, this becomes a formidable task. Thus, a theoretically attractive approach to ecological inference is a NCAR strategy where $g(X, W | \theta, \gamma)$ is estimated with a non/semiparametric model. In what follows, we will discuss this modeling strategy and compare its performance with that of other methods using simulated and empirical data.

3 Data Augmentation Approach to Ecological Inference

In this section, we introduce our modeling strategies based on the theoretical framework developed in Section 2. Since coarse data are a particular form of missing data, it is natural to consider the method of data augmentation, which has been widely used when analyzing various missing data problems (e.g. Tanner and Wong, 1987; van Dyk and Meng, 2001). The basic idea of data augmentation is to repeat the following two steps until a satisfactory degree of convergence is achieved; (1) predict the values of missing data based on the complete-data model, and (2) estimate the complete-data model parameters based on the augmented data. In the context of coarse data, we refine the coarse data to predict the unobserved values of W , and then estimate θ based on the imputed values of W .

In addition to its intuitive appeal, the method of data augmentation can also incorporate individual-level information whenever such additional information is available. For example, exit polls or other types of survey data in some geographical units might be available to researchers who study racial voting through ecological inference. Many have pointed out that a small amount of individual-level information can lead to dramatic improvement of resulting ecological inference (e.g., Wakefield, 2004; Imai and Lu, 2004). In the following, we present our modeling strategies

under each of the three assumptions; CAR, CCAR, and NCAR.

3.1 Modeling under CAR

Imai and Lu (2004) proposed a parametric model based on the CAR assumption (see also Wakefield, 2004). In particular, they make the following distributional assumption for the logit transformation of W_{1i} and W_{2i} , which we denote by $W_{1i}^* = \log\left(\frac{W_{1i}}{1-W_{1i}}\right)$ and $W_{2i}^* = \log\left(\frac{W_{2i}}{1-W_{2i}}\right)$,

$$(W_{1i}^*, W_{2i}^*) \stackrel{\text{indep.}}{\sim} \mathcal{N}(\mu, \Sigma). \quad (12)$$

Similar to the model of King (1997), this model allows W_{1i} and W_{2i} to be correlated with each other (through their logit transformations). This implies the complete-data likelihood function of $W = (W_1, W_2)$,

$$f(W \mid \mu, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|} W_1 W_2 (1 - W_1)(1 - W_2)} \exp\left[-\frac{1}{2} \{\text{logit}(W) - \mu\}^\top \Sigma^{-1} \{\text{logit}(W) - \mu\}\right]. \quad (13)$$

Together with $r(Y \mid W, X)$ of equation (4), we can specify the observed likelihood function $L_{obs}(\mu, \Sigma \mid Y)$ of equation (8) under CAR. To compute the maximum likelihood estimates of μ and Σ , we use the *EM* algorithm of Dempster *et al.* (1977). Furthermore, the Supplemented *EM* (*SEM*) algorithm of Meng and Rubin (1991) can be used to estimate the asymptotic variance-covariance matrix.

An important benefit of using these algorithms is that we can quantify the effect of missing information on parameter estimation by applying the missing information principle of Orchard and Woodbury (1972); i.e., states *observed information = complete information - missing information*. Specifically, we write $I_o = I_{oc} - I_{om}$ where I_o denotes the information matrix based on the observed data, I_{oc} is the expected information matrix based on the complete data, and I_{om} represents the

missing information. The *SEM* algorithm allows us to estimate I_o and I_{oc} (Meng and Rubin, 1991), and based on these estimates, we can calculate the estimated amount of missing information I_{om} . The estimated fraction of missing information for each parameter can be computed as the ratio between the corresponding diagonal element of I_{om} and I_{oc} . A large fraction of missing information suggests that the parameter estimate is heavily affected by data coarsening and less reliable. We believe that this consideration is critical in the context of ecological inference because it identifies the amount of information loss due to aggregation and directly relates to the model assessment.

Imai and Lu (2004) further show that this parametric model can be estimated using a fully Bayesian model by specifying a prior distribution for (μ, Σ) . They use a conjugate prior and implement a Markov chain Monte Carlo (MCMC) algorithm to fit the model. Imai and Lu (2004) also demonstrate that this parametric Bayesian model can be extended to a nonparametric Bayesian model based on the Dirichlet process prior (Ferguson, 1973; Escobar and West, 1995), which relaxes the distributional assumption concerning W . They find that the nonparametric model tends to outperform various parametric models both in terms of in-sample and population inferences.

3.2 Modeling under CCAR

The parametric model defined above can be modified to fit the assumption of CCAR. This extension can be accomplished by specifying the conditional distribution of W_i given a $(k \times 1)$ vector of the observed covariates Z_i , which does not include X_i . In particular, we assume that $W_i^* = (W_{1i}^*, W_{2i}^*)$ follows a bivariate normal distribution with a mean vector $B^\top Z_i$ where B is a $(k \times 2)$ matrix of coefficients including intercepts (see also King, 1997; King *et al.*, 1999). Formally,

this can be written as the following multivariate regression model,

$$W_i^* | B, \Sigma, Z_i \stackrel{\text{indep.}}{\sim} \mathcal{N}(B^\top Z_i, \Sigma), \quad (14)$$

where Σ is the (2×2) positive definite population covariance matrix. Equation (14) defines $f(W | Z, \theta)$ where $\theta = (B, \Sigma)$, and together with $r(Y | W, X, \theta)$, this specifies the observed-data likelihood of equation (10).

The Expectation-Conditional Maximization (*ECM*) and Supplemented *ECM* algorithms can be employed to obtain the maximum likelihood estimates of B and Σ (Meng and Rubin, 1993; van Dyk *et al.*, 1995). Alternatively, a fully Bayesian model can be specified by placing the normal/inverse-Wishart conjugate prior distribution on (B, Σ) . The conditional prior distribution of B given Σ is $B | \Sigma \sim \mathcal{MN}(B_0, A_0^{-1}, \Sigma)$ where $\mathcal{MN}()$ is a matrix-variate normal distribution, B_0 is a $k \times 2$ matrix of the prior mean of B , A_0 is a $(k \times k)$ scale matrix, Σ is the covariance matrix. If we vectorize B into a $(2k, 1)$ vector, then $\text{vec}(B) | \Sigma \sim \mathcal{N}(\text{vec}(B_0), A_0^{-1} \otimes \Sigma)$. Finally, the prior distribution of Σ is $\Sigma \sim \text{InvWish}(\nu_0, S_0^{-1})$ where ν_0 is the degrees of freedom parameter, and S_0 is a 2×2 positive definite prior scale matrix for Σ .

3.3 Modeling under NCAR

We further extend the model to the situation where the CAR assumption does not hold (i.e., NCAR). To formulate a parametric model, we model $(W_{1i}^*, W_{2i}^*, X_i^*)$ jointly using the following trivariate normal distribution,

$$(W_{1i}^*, W_{2i}^*, X_i^*) \stackrel{\text{indep.}}{\sim} \mathcal{N}(\eta, \Phi), \quad (15)$$

where $X_i^* = \log\left(\frac{X_i}{1-X_i}\right)$, η is the (3×1) vector of means, and Φ is the (3×3) positive definite covariance matrix. This models allows pair-wise correlations among W_{1i} , W_{2i} , and X_i through

their logit transformations and specifies $g(X, W | \theta, \gamma)$ of the observed likelihood of equation (11).

To obtain the maximum likelihood estimates of the parameters, one can employ the *EM* algorithms by evaluating the conditional expectation of W_i^* given X_i^* as well as the parameter values from the previous iteration. Alternatively, we can also formulate a fully Bayesian model by placing the following conjugate prior distribution on (η, Φ) ; i.e., $\eta | \Phi \sim \mathcal{N}(\eta_0, \Phi/\tau_0^2)$, 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 proper, ν_0 needs to be greater than 3. A Gibbs sampling algorithm can be used to fit this Bayesian model.

We can follow the strategy of Imai and Lu (2004) and relax the distributional assumption concerning $g(X, W | \theta, \gamma)$ under the NCAR assumption. We model the triplet $(W_{1i}^*, W_{2i}^*, X_i^*)$ as a mixture of normals by applying a Dirichlet process prior distribution to the model parameters. Such a Bayesian nonparametric model is given by,

$$(W_{1i}^*, W_{2i}^*, X_i^*) | \eta_i, \Phi_i \stackrel{\text{indep.}}{\sim} \mathcal{N}(\eta_i, \Phi_i), \quad (16)$$

$$\eta_i, \Phi_i | G \stackrel{\text{i.i.d.}}{\sim} G, \quad (17)$$

$$G | \alpha \sim \mathcal{D}(G_0, \alpha) \quad (18)$$

$$\alpha \sim \mathcal{G}(a_0, b_0) \quad (19)$$

where G denotes the random distribution function of (η_i, Φ_i) and α is the concentration parameter of the Dirichlet distribution. Under the base distribution $G_0 = E(G)$, (η_i, Φ_i) is distributed as a trivariate normal/inverse-Wishart distribution, $\eta_i | \Phi_i \sim \mathcal{N}(\eta_0, \Phi_i/\tau_0^2)$, and $\Phi_i \sim \text{InvWish}(\nu_0, S_0^{-1})$. When α is large, G is close to G_0 and hence not different from the parametric model, while smaller value of α puts most probability mass on a few partitions of the

parameter space hence allows to model more flexible shapes of distributions. This model is a three-dimension extension of the two-dimensional model as described in Imai and Lu (2004). The Gibbs sampling algorithm can be used to fit this model.

4 Simulation and Empirical Studies

In this section, we study the performance of the models proposed in Section 3 by analyzing simulated and empirical data. First, we illustrate that even when the CAR assumption holds (i.e., no contextual effect), the information loss due to aggregation process still affects the model performance. Second, we conduct a simulation study to show that if we properly control for the covariates, the ecological inference based on the CCAR assumption yields reasonable results. However, the misspecification of the model can yield biased estimates. Lastly, we reanalyze a classical ecological inference problem about black illiteracy rates in 1910 (Robinson, 1950; King, 1997). Using this example, we examine the performance of both parametric and nonparametric NCAR models when the CAR assumption is violated.

4.1 Quantifying Aggregation Effects

We analyze the voter registration data from 144 counties of North Carolina and South Carolina. This is a subset of the voter registration data of the southern states in the United States that was used by King (1997). For each county, X_i represents the proportion of black voters, Y_i denotes the registration rate, W_{1i} and W_{2i} are the registration rates of black and white voters. Through the analysis of this dataset, we show that *EM* algorithms can quantify the missing information due to aggregation. The estimated fraction of missing information we obtain quantifies the degree

to which aggregation effects influence the parameter estimation. We also show that by including some individual level information, the influence of missing information can be mitigated to some extent.

In this example, W_1 and W_2 are actually observed, and hence, we know that the data closely fit the CAR assumption. The proportion of the black voters X is only slightly correlated with the black registration rate W_1 and with the white registration rate W_2 ; the correlation is approximately 0.2 in both cases. In addition, the average bounds length of W_1 is 0.83, while that of W_2 is only 0.32. This indicates that there is less information about W_1 . The sample mean of the black registration rate is 0.5 and that of white registration rate is 0.78. Since we know the true values of (W_1, W_2) in this example, we substitute different proportions of the coarsened observations by their corresponding true values in order to examine how the impact of missing information varies. In particular, we consider five scenarios in which the proportion of the true values varies from 0 percent to 20 percent.

Based on the parametric model under the CAR assumption, we first estimate the mean, variance and covariance $(\mu_1, \mu_2, \sigma_{11}, \sigma_{22}, \sigma_{12})$ of the bivariate normal complete-data likelihood and their asymptotic variance using the *EM* and *SEM* algorithms. We then compute the estimated fraction of missing information for each parameter. We do this for each scenario and assess the impact of missing information on parameter estimation.

Figure 1 illustrates how the estimated fraction of missing information decreases as we increase the amount of individual level data. Initially, without any individual level information, the estimated fraction of missing information is high for all the parameter estimates. Especially for the parameters related to W_1 , $(\mu_1, \sigma_{11}, \sigma_{12})$, the corresponding estimated fractions of missing information are as high as 90%. As the amount of true values increases, the estimated fraction of missing

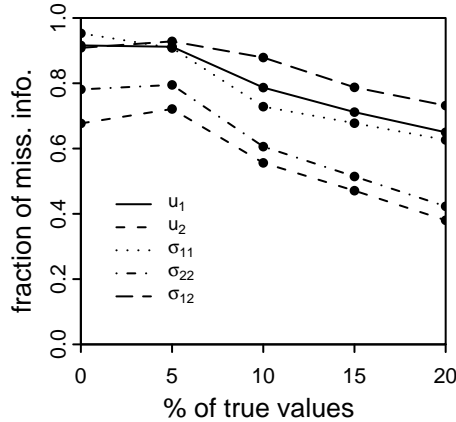


Figure 1: Estimated Fraction of Missing Information for Each Parameter when Different Amount of True Individual Data is Given. The vertical axis indicates the fraction of missing information, while the horizontal axis indicates the amount of individual-level data that are included. Each line connects the estimated fraction of missing information for each parameter under different scenarios.

information declines.

A large amount of missing information makes the parameter estimation inaccurate. For example, without additional individual level data, the estimated fraction of missing information for μ_1 is 0.92, and 0.68 for μ_2 . While $\hat{\mu}_1$ is 0.66 with standard error 0.33, and $\hat{\mu}_2$ is 0.80 with standard error 0.19. When compared to the true sample means, $\hat{\mu}_2$ appears to be a better estimate. When 20 percent of individual level data is supplied, the fraction of missing information associated with μ_1 is reduced to 65%, and $\hat{\mu}_1$ is 0.51 with standard error 0.16 which appears to be a more accurate estimate.

4.2 Modeling Contextual Effects

In this section, we investigate the possibility of correcting the contextual effect by controlling other covariates. To do this, we conduct a simulation study considering a situation where X and W are

correlated through a third variable Z . Specifically, X and W are conditionally independent given Z .

4.2.1 A Simulation Study.

To avoid other confounding issues, we simulate a dataset 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 $(W_{1i}, W_{2i}, X_i, Z_i)$, denoted by $(W_{1i}^*, W_{2i}^*, X_i^*, Z_i^*)$ 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 $W_i^* = BZ_i^* + \epsilon_{1i}$, 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 zero. ϵ_{1i} 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, $X_i^* = 2Z_i^* + 0.5Z_i^{*2} + \epsilon_{2i}$, where ϵ_{2i} is an independent draw from a univariate normal distribution with mean zero 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 , for W_2 is 0.4 , suggesting that W_1 is more coarsened than W_2 . Finally, the

| Model | Bias | | RMSE | |
|------------------------|--------|--------|-------|-------|
| | 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 |

Table 2: In-sample Predictive Performance of the Parametric Models When X and W Are Independent Given Z . The bias for W_j is calculated as $\sum_{i=1}^n (\widehat{W}_{ji} - W_{ji})/n$ for $j = 1, 2$, where \widehat{W}_{ji} denotes the in-sample predictions of W_{ji} , and W_{ji} is the true value. Similarly, the root mean squared error (RMSE) is defined as $\sqrt{\sum_{i=1}^n (\widehat{W}_{ji} - W_{ji})^2/n}$.

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 Bayesian models we introduced in Sections 3.2 and 3.3, we adopt diffuse prior distributions. In particular, for the parametric CAR model, we use the same prior specifications as in Imai and Lu (2004). For the parametric CCAR model, the prior parameters are $B_0 = \mathbf{0}$, $A_0 = I_2$, $\nu_0 = 7$, and $S_0 = 10I_2$. While for the parametric NCAR model, our choices of diffuse prior distribution is defined by $\eta_0 = \mathbf{0}$, $\tau_0 = 2$, $\nu_0 = 5$, and $S_0 = 13I_2$.

Table 2 presents the bias and root mean square error (RMSE) of the *in-sample* predictions based for each parametric model. As we expected, when the correct covariate is controlled, CCAR model yields the smallest bias and root mean square error. 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 which 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 predictions of W_2 based on this model still improves substantially comparing to those based on the CAR model and the misspecified CCAR models.

4.2.2 An Empirical Example: Race and Literacy

Since W_1 and W_2 are not observed, applied researchers often cannot rely on the CAR assumption. Because of potential misspecification, the CCAR model may not be a realistic alternative. In some cases, researchers may not possess the knowledge about which variables to condition on. In other situations, such variables may not be available. 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 no necessary correspondence between aggregate and individual-level correlations. The original study was 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 dataset, there are 1,040 counties. The dataset 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 illiteracy 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 dataset, the black literacy rate is negatively correlated with the percentage of black population X (the sample correlation is -0.51), while the white literacy rate is 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 4.2. For the nonparametric CAR and NCAR models, the corresponding

| | 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 |
| Ecological regression | -0.072 | 0.016 | 0.128 | 0.059 |

Table 3: In-sample Predictive Performance of Various Models When X and W are Correlated. The bias for W_j is calculated as $\sum_{i=1}^n (\widehat{W}_{ji} - W_{ji})/n$ for $j = 1, 2$, where \widehat{W}_{ji} denotes the in-sample predictions of W_{ji} , and W_{ji} is the true value. Similarly, the root mean squared error (RMSE) is defined as $\sqrt{\sum_{i=1}^n (\widehat{W}_{ji} - W_{ji})^2/n}$.

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 , i.e., Gamma(1, 0.1) as in Imai and Lu (2004).

Table 3 summarizes the in-sample predictive performance. The results based on the King's EI models and the ecological regression are also presented. The NCAR models clearly outperform the other models. When compared to the other models, the in-sample predictions of the NCAR models are closer to being unbiased and have relatively small root mean squared errors. Indeed, the correlation between the in-sample predictions of W_1 and X is approximately -0.51 for the parametric NCAR model and -0.56 for the nonparametric NCAR model which are both very close to the correlation estimate based on the true values. This indicates the advantage of controlling the contextual effects via the NCAR models. Although we incorporated X as a covariate in the King's extended EI model, in this particular example this strategy does not appear to be as effective as the NCAR models. When the contextual effect is ignored as in the CAR models, the in-sample

| True value | CAR Models | | NCAR Models | | King’s EI Models | | Ecological Regression |
|------------|------------|----------|-------------|----------|------------------|--------------|-----------------------|
| | Para. | Nonpara. | Para. | Nonpara. | With covariate | No covariate | |
| -0.339 | -0.414 | -0.410 | -0.359 | -0.341 | -0.411 | -0.412 | -0.401 |

Table 4: Estimated Individual Correlations Based on Different Models. 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 .

predictions of W_{1i} are severely biased.

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 4. 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 Concluding Remarks

The incomplete data framework and the data augmentation approach proposed in this paper together offer various modeling strategies in ecological inference. Given this toolkit, applied researchers can decide how to model the distribution of the unknowns $f(W | \theta)$ and the stochastic coarsening process $h(X | W, \gamma)$. In addition, many existing methods can be revisited under the proposed framework. For example, the method of bounds that uses the midpoints of bounds as in-sample estimates is equivalent to assuming a uniform distribution of W and coarsened at random (CAR). Similarly, King’s basic EI model adopts a truncated bivariate normal distribution of W while maintaining the CAR assumption. His extended model conditions on covariates by

assuming CCAR. In the presence of the spatial effects, one can model $f(W | \theta)$ using hierarchical structure with spatial correlation.

Furthermore, we can formally understand the difficulties of ecological inference by formulating ecological inference problem as an incomplete data problem. Under the statistical framework of coarse data, three key issues – distributional, contextual and aggregation effects – can be formally identified and modeled. By quantifying the amount of missing information, we see that in some cases the information loss due to aggregation is so severe that it greatly affects the precision of parameter estimation even when distributional and contextual effects are not present.

The bias in ecological inference due to contextual effects is known to be difficult to cope with. We have shown that this bias can be significantly reduced by jointly modeling the complete data likelihood and the stochastic coarsening process (NCAR) or by properly conditioning on the variables that determine the coarsening process (CCAR). However, if one adopts the latter strategy, the inclusion of spurious covariates and the violation of other functional form assumptions can lead to misleading results. When researchers do not possess strong knowledge about the model specification, the NCAR models are preferable.

Finally, another challenge of ecological inference is the specification of the distributional assumption about partially observed variables. Since the distributional assumptions of the incomplete data and the coarsening process are not directly verifiable, we recommend that researchers avoid strong parametric assumptions and use more flexible models such as the Bayesian nonparametric models proposed in this paper. Using simulation studies, Imai and Lu (2004) show that the nonparametric model under the CAR assumption is an effective estimation tool for both the in-sample predictions and population inference. In our simulation and empirical studies, we find that the nonparametric NCAR model performs slightly better than the parametric NCAR model.

References

- Achen, C. H. and Shively, W. P. (1995). *Cross-Level Inference*. University of Chicago Press, Chicago.
- Alesina, A. and Rosenthal, H. (1995). *Partisan Politics, Divided Government, and the Economy*. Cambridge University Press, New York.
- Burden, B. C. and Kimball, D. C. (1998). A new approach to the study of ticket splitting. *American Political Science Review* **92**, 3, 533–544.
- Cho, W. K. T. (1998). If the assumptions fits...:a comment on the King ecological inference solution. *Political Analysis* **7**, 143–163.
- Darby, S., Deo, H., and Doll, R. (2001). A parallel analysis of individual and ecological data on residential radon and lung cancer in south-west england. *Journal of the Royal Statistical Society, Series A* **164**, 193–203.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **39**, 1–37.
- Duncan, O. D. and Davis, B. (1953). An alternative to ecological correlation. *American Sociological Review* **18**, 665–666.
- Durkheim, E. (1897). *Le Suicide, English translation by J. A. Spalding in 1951*. Free Press, Toronto, Canada.
- Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.
- Fiorina, M. P. (1996). *Divided Government*. Allyn and Bacon, Needham Heights, MA.
- Freedman, D. A. (1999). Ecological inference and the ecological fallacy. In N. Smelser and P. Baltes, eds., *International Encyclopedia of the Social and Behavioral Sciences*, vol. 6, 4027–4030. Elsevier.
- Freedman, D. A., Klein, S. P., Sacks, J., Smyth, C. A., and Everett, C. G. (1991). Ecological regression and voting rights (with discussion). *Evaluation Review* **15**, 673–816.
- Freedman, D. A., Ostland, M., Roberts, M. R., and Klein, S. P. (1998). “Review of ‘A Solution to the Ecological Inference Problem’ ”. *Journal of the American Statistical Association* **93**, 1518–1522.
- Goldsmith, J. R. (1999). The residential random-lung cancer association in u.s. counties: A commentary. *Health Physics* **76**, 553–557.
- Goodman, L. (1953). Ecological regressions and behavior of individuals. *American Sociological Review* **18**, 663–666.
- Goodman, L. A. (1959). Some alternatives to ecological correlation. *The American Journal of Sociology* **64**, 610–624.
- Grofman, B. (1991). Statistics without substance: A critique of Freedman et al. and Clark and Morrison. *Evaluation Review* **15**, 6, 746–769.
- Hajnal, Z. L., Gerber, E. R., and Louch, H. (2002). Minorities and direct legislation: Evidence from california ballot proposition elections. *The Journal of Politics* **64**, 154–177.

- Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *The Annals of Statistics* **19**, 2244–2253.
- Imai, K. and Lu, Y. (2004). Parametric and nonparametric Bayesian models for ecological inference in 2×2 tables. *Proceedings of the American Statistical Association* available at <http://imai.princeton.edu/research/einonpar.html>.
- Imai, K. and Lu, Y. (2005). `eco`: R package for fitting bayesian models of ecological inference in 2×2 tables. available at The Comprehensive R Archive Network (CRAN). <http://cran.r-project.org>.
- Johnston, R. and Pattie, C. (2000). Ecological inference and entropy-maximizing: an alternative estimation procedure for split-ticket voting. *Political Analysis* **8**, 333–345.
- King, G. (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press, Princeton, NJ.
- King, G. (1999). Comment on “Review of ‘A Solution to the Ecological Inference Problem’”. *Journal of the American Statistical Association* **94**, 352–355.
- King, G., Rosen, O., Tanner, M., and Wagner, A. F. (2004a). Ordinary voting behavior in the extraordinary election of adolf hitler. *Typescript, Department of Government, Harvard University* .
- King, G., Rosen, O., and Tanner, M. A. (1999). Binomial-beta hierarchical models for ecological inference. *Sociological Methods & Research* **28**, 61–90.
- King, G., Rosen, O., and Tanner, M. A., eds. (2004b). *Ecological Inferece: New Methodological Strategies*. Cambridge University Press.

- Lichtman, A. J. (1991). Passing the test: Ecological regression analysis in the Los Angeles county case and beyond. *Evaluation Review* **15**, 6, 770–799.
- Little, R. J. and Rubin, D. B. (1987). *Statistical Analysis With Missing Data*. John Wiley & Sons, New York.
- Lohmoller, J.-B., Falter, J., Link, A., and de Rijke, J. (1985). *Measuring the Unmeasurable*, chap. Unemployment and the Rise of National Socialism: Contradicting Results From Different Regional Aggregations, 357–370. The Hague:Martinus Nijhoff.
- Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association* **86**, 899–909.
- Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**, 267–278.
- Neeleman, J. and Lewis, G. (1999). Suicide, religion, and socioeconomic conditions: An ecological study in 26 countries, 1990. *Journal of Epidemiology and Community Health* **53**, 204–210.
- Neyman, J. and Scott, E. L. (1948). Consistent estimation from partially consistent observations. *Econometrica* **16**, 1–32.
- O’Loughlin, J. (2000). Can King’s ecological inference method answer a social scientific puzzle: Who voted for the Nazi party in Weimar Germany? *Annals of the Association of American Geographers* **90**, 592–601.
- Orchard, T. and Woodbury, M. A. (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**, 351–357.
- Rosen, O., Jiang, W., King, G., and Tanner, M. A. (2001). Bayesian and frequentist inference for ecological inference: The $R \times C$ case. *Statistica Neerlandica* **55**, 2, 134–156.
- Rubin, D. B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association* **69**, 467–474.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.
- Shaw, D. R. (1997). Estimating racially polarized voting: A view from the states. *Political Research Quarterly* **50**, 49–74.
- Snow, J. (1855). *On the Mode of Communication of Cholera*(Reprinted in 1965). Hafner, London.
- Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528–550.
- van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation (with discussions). *The Journal of Computational and Graphical Statistics* **10**, 1–111.
- van Dyk, D. A., Meng, X.-L., and Rubin, D. B. (1995). Maximum likelihood estimation via the ECM algorithm: Computing the asymptotic variance. *Statistica Sinica* **5**, 55–75.
- Wakefield, J. (2004). Ecological inference for 2×2 tables (with discussion). *Journal of the Royal Statistical Society, Series A* **167**, 385–445.