INTRODUCTION

Information in Ecological Inference: An Introduction

Gary King, Ori Rosen, and Martin A. Tanner

Researchers in a diverse variety of fields often need to know about individual-level behavior and are not able to collect it directly. In these situations, where survey research or other means of individual-level data collection are infeasible, ecological inference is the best and often the only hope of making progress. Ecological inference is the process of extracting clues about individual behavior from information reported at the group or aggregate level.

For example, sociologists and historians try to learn who voted for the Nazi party in Weimar Germany, where thoughts of survey research are seven decades too late. Marketing researchers study the effects of advertising on the purchasing behavior of individuals, where only zip-code-level purchasing and demographic information are available. Political scientists and politicians study precinct-level electoral data and U.S. Census demographic data to learn about the success of candidate appeals with different voter groups in numerous small areal units where surveys have been infeasible (for cost or confidentiality reasons). To determine whether the U.S. Voting Rights Act can be applied in redistricting cases, expert witnesses, attorneys, judges, and government officials must infer whether African Americans and other minority groups vote differently from whites, even though the secret ballot hinders the process and surveys in racially polarized contexts are known to be of little value.

In these and numerous other fields of inquiry, scholars have no choice but to make ecological inferences. Fortunately for them, we have witnessed an explosion of statistical research into this problem in the last five years – both in substantive applications and in methodological innovations. In applications, the methods introduced by Duncan and Davis (1953) and by Goodman (1953) accounted for almost every use of ecological inference in any field for fifty years, but this stasis changed when King (1997) offered a model that combined and extended the approaches taken in these earlier works. His method now seems to dominate substantive research in academia, in private industry, and in voting rights litigation, where it was used in most American states in the redistricting period that followed the 2000 Census. The number and diversity of substantive application areas of ecological inference has soared recently as well. The speed of development of statistical research on ecological inference has paralleled the progress in applications, too, and in the last five years we have seen numerous new models, innovative methods, and novel computation schemes. This book offers a snapshot of some of the research at the cutting edge of this field in the hope of spurring statistical researchers to push out the frontiers and applied researchers to choose from a wider range of approaches.

Ecological inference is an especially difficult special case of statistical inference. The difficulty comes because some information is generally lost in the process of aggregation, and that information is sometimes systematically related to the quantities of interest. Thus, progress

in this field has usually come from discovering new sources of information or inventing better ways of harvesting existing information and using it to improve our inferences about individual-level behavior. This book is organized around these sources of information and methods for their extraction. We begin this overview chapter in Section 0.1 by very briefly summarizing some relevant prior research, on which the authors in this volume build. This section also serves to introduce the notation used, when convenient, in the rest of the book. Section 0.2 then summarizes the subsequent chapters.

0.1 NOTATION AND BACKGROUND

0.1.1 The Ecological Inference Problem

For expository purposes, we discuss only an important but simple special case of ecological inference, and adopt the running example and notation from King (1997: Chapter 2). The basic problem has two observed variables (T_i and X_i) and two unobserved quantities of interest (β_i^b and β_i^w) for each of p observations. Observations represent aggregate units, such as geographical areas, and each individual-level variable within these units is dichotomous.

To be more specific, in Table 0.1, we observe for each electoral precinct i ($i = 1, \ldots, p$) the fractions of voting age people who turn out to vote (T_i) and who are black (X_i), along with the number of voting age people (N_i). The quantities of interest, which remain unobserved because of the secret ballot, are the proportions of blacks who vote (β_i^b) and whites who vote (β_i^w). The proportions β_i^b and β_i^w are not observed because T_i and X_i are from different data sources (electoral results and census data, respectively) and record linkage is impossible (and illegal), and so the cross-tabulation cannot be computed.

Also of interest are the district-wide fractions of blacks and whites who vote, which are respectively

$$B^b = \frac{\sum_{i=1}^{p} N_i X_i \beta_i^b}{\sum_{i=1}^{p} N_i X_i} \tag{0.1}$$

and

$$B^w = \frac{\sum_{i=1}^{p} N_i (1 - X_i) \beta_i^w}{\sum_{i=1}^{p} N_i (1 - X_i)}. \tag{0.2}$$

These are weighted averages of the corresponding precinct-level quantities. Some methods aim to estimate only B^b and B^w without giving estimates of β_i^b and β_i^w for all i.

0.1.2 Deterministic and Statistical Approaches

The ecological inference literature before King (1997) was bifurcated between supporters of the method of bounds, originally proposed by Duncan and Davis (1953), and supporters of statistical approaches, proposed even before Ogburn and Goltra (1919), but first formalized into a coherent statistical model by Goodman (1953, 1959).[1] Although Goodman and

[1] For the historians of science among us: despite the fact that these two monumental articles were written by two colleagues and friends in the same year and in the same department and university (the Department of Sociology at the University of Chicago), the principals did not discuss their work prior to completion. Even judging by today's standards, nearly a half-century after their publication, the articles are models of clarity and creativity.

Table 0.1 Notation for precinct i

Race of voting age person	Voting decision: Vote	Voting decision: No vote	
Black	β_i^b	$1-\beta_i^b$	X_i
White	β_i^w	$1-\beta_i^w$	$1-X_i$
	T_i	$1-T_i$	

Note: The goal is to estimate the quantities of interest, β_i^b (the fraction of blacks who vote) and β_i^w (the fraction of whites who vote), from the aggregate variables X_i (the fraction of voting age people who are black) and T_i (the fraction of people who vote), along with N_i (the known number of voting age people).

Duncan and Davis moved on to other interests following their seminal contributions, most of the ecological inference literature in the five decades since 1953 was an ongoing war between supporters of these two key approaches, often without the usual academic decorum.

0.1.2.1 Extracting Deterministic Information: The Method of Bounds

The purpose of the method of bounds and its generalizations is to extract deterministic information, known with certainty, about the quantities of interest.

The intuition behind these quantities is simple. For example, if a precinct contained 150 African-Americans and 87 people in the precinct voted, then how many of the 150 African-Americans actually cast their ballot? We do not know exactly, but bounds on the answer are easy to obtain: in this case, the answer must lie between 0 and 87. Indeed, conditional only on the data being correct, $[0, 87]$ is a 100% confidence interval. Intervals like this are sometimes narrow enough to provide meaningful inferences, and sometimes they are too wide, but the ability to provide (nontrivial) 100% confidence intervals in even some situations is quite rare in any statistical field.

In general, before seeing any data, the unknown parameters β_i^b and β_i^w are each bounded on the unit interval. Once we observe T_i and X_i, they are bounded more narrowly, as

$$\begin{aligned} \beta_i^b &\in \left[\max\left(0, \frac{T_i-(1-X_i)}{X_i}\right), \ \min\left(\frac{T_i}{X_i}, 1\right)\right], \\ \beta_i^w &\in \left[\max\left(0, \frac{T_i-X_i}{1-X_i}\right), \ \min\left(\frac{T_i}{1-X_i}, 1\right)\right]. \end{aligned} \tag{0.3}$$

Deterministic bounds on the district-level quantities B^b and B^w are weighted averages of these precinct-level bounds.

These expressions indicate that the parameters in each case fall within these deterministic bounds with certainty, and in practice they are almost always narrower than $[0, 1]$. Whether they are narrow enough in any one application depends on the nature of the data.

0.1.2.2 Extracting Statistical Information: Goodman's Regression

Leo Goodman's (1953, 1959) approach is very different from Duncan and Davis's. He looked at the same data and focused on the statistical information. His approach examines variation in the marginals (X_i and T_i) over the precincts to attempt to reason back to the district-wide fractions of blacks and whites who vote, B^b and B^w. The outlines of this approach, and the problems with it, have been known at least since Ogburn and Goltra (1919). For example, if in precincts with large proportions of black citizens we observe that many people do not vote, then it may seem reasonable to infer that blacks turn out at rates lower than whites. Indeed it often is reasonable, but not always. The problem is that it could instead be the case that the whites who happen to live in heavily black precincts are the ones who vote less frequently, yielding the opposite ecological inference with respect to the individual-level truth.

What Goodman accomplished was to formalize the logic of the approach in a simple regression model, and to give the conditions under which estimates from such a model are unbiased. To see this, note first that the accounting identity

$$T_i = X_i\beta_i^b + (1 - X_i)\beta_i^w \tag{0.4}$$

holds exactly. Goodman showed that a regression of T_i on X_i and $1 - X_i$ with no constant term could be used to estimate B^b and B^w, respectively. The key assumption necessary for unbiasedness that Goodman identified is that the parameters and X_i are uncorrelated: $\text{Cov}(\beta_i^b, X_i) = \text{Cov}(\beta_i^w, X_i) = 0$. In the example, the assumption is that blacks vote in the same proportions in homogeneously black areas as in more integrated areas.[2] Obviously, this is true sometimes and it is false at other times.

As Goodman recognized, when this key assumption does not hold, estimates from the model will be biased. Indeed, they can be very biased, outside the deterministic bounds, and even outside the unit interval. Goodman's technique has been used extensively in the last half-century, and impossible estimates occur with considerable frequency (some say in a majority of real applications; see Achen and Shively, 1995).

0.1.3 Extracting Both Deterministic and Statistical Information: King's EI Approach

From 1953 until 1997, the only two approaches used widely in practice were the method of bounds and Goodman's regression. King's (1997) idea was that the insights from these two conflicting literatures in fact do not conflict with each other; the sources of information are largely distinct and can be combined to improve inference overall and synergistically. The idea is to combine the information from the bounds, applied to both quantities of interest for each and every precinct, with a statistical approach for extracting information within the bounds. The amount of information in the bounds depends on the data set, but for many data sets it can be considerable. For example, if precincts are spread uniformly over a scatterplot of X_i by T_i, the average bounds on β_i^b and β_i^w are narrowed from $[0, 1]$ to less than half of that range – hence eliminating half of the ecological inference problem with certainty. This additional information also helps make the statistical portion of the model far less sensitive to assumptions than previous statistical methods that exclude the information from the bounds.

To illustrate these points, we first present all the information available without making any assumptions, thus extending the bounds approach as far as possible. As a starting point, the

[2] King (1997: Chapter 3) showed that Goodman's assumption was necessary but not sufficient. To have unbiasedness, it must also be true that the parameters and N_i are uncorrelated.

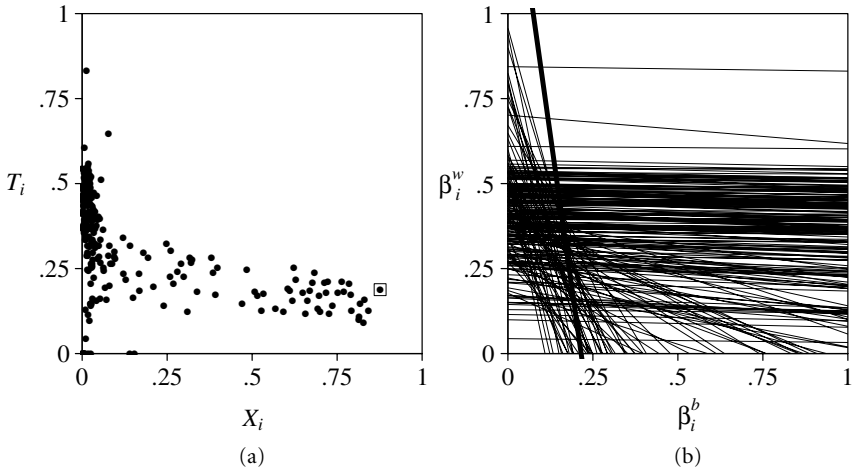

Figure 0.1. Two views of the same data: (a) a scatterplot of the observables, X_i by T_i; (b) this same information as a tomography plot of the quantities of interest, β_i^b by β_i^w. Each precinct i that appears as a point in (a) appears instead as a line (because of information lost due to aggregation) in (b). For example, precinct 52 appears as the dot with a little square around it in (a), and as the dark line in (b). The data are from King (1997: Figures 5.1, 5.5).

graph in Figure 0.1a provides a scatterplot of a sample data set as observed, X_i horizontally by T_i vertically. Each point in this plot corresponds to one precinct, for which we would like to estimate the two unknowns. We display the unknowns in part (b) of the same figure; any point in that graph portrays values of the two unknowns, β_i^b (plotted horizontally) and β_i^w (vertically). Ecological inference involves locating, for each precinct, the one point in this unit square corresponding to the true values of β_i^b and β_i^w, since values outside the square are logically impossible.

To map the knowns onto the unknowns, King began with Goodman's accounting identity from Equation 0.4. From this equation, which holds exactly, we solve for one unknown in terms of the other:

$$\beta_i^w = \left(\frac{T_i}{1 - X_i}\right) - \left(\frac{X_i}{1 - X_i}\right)\beta_i^b, \tag{0.5}$$

which shows that β_i^w is a *linear* function of β_i^b, where the intercept and slope are known (since they are functions of the data, X_i and T_i).

King then maps the knowns from Figure 0.1a onto Figure 0.1b by using the linear relationship in Equation 0.5. A key point is that each dot in (a) can be expressed, without assumptions or loss of information, as what King called a "tomography" line within the unit square in (b).[3] It is precisely the information lost due to aggregation that causes us to have to plot an entire line (on which the true point must fall) rather than the goal of one point for each precinct in Figure 0.1b. In fact, the information lost is equivalent to having a graph of the (β_i^b, β_i^w) points but having the ink smear, making the points into lines and partly but not entirely obscuring the correct positions of the points.

[3] King also showed that the ecological inference problem is mathematically equivalent to the ill-posed "tomography" problem of many medical imaging procedures (such as CAT and PET scans), where one attempts to reconstruct the inside of an object by passing X-rays through it and gathering information only from the outside. Because the line sketched out by an X-ray is closely analogous to Equation 0.5, King called the latter a *tomography line* and the corresponding graph a *tomography graph.*

What does a tomography line tell us? Before we know anything, we know that the true (β_i^b, β_i^w) point must lie somewhere within the unit square. After X_i and T_i are observed for a precinct, we also know that the true point must fall on a specific line represented by Equation 0.5 and appearing in the tomography plot in Figure 0.1. In many cases narrowing the region to be searched for the true point from the entire square to the one line in the square can provide a significant amount of information. To see this, consider the point enclosed in a box in Figure 0.1a, and the corresponding dark line in Figure 0.1b. This precinct, number 52, has observed values of $X_{52} = 0.88$ and $T_{52} = 0.19$. As a result, substituting into Equation 0.5 gives $\beta_i^w = 1.58 - 7.33\beta_i^b$, which when plotted then appears as the dark line in (b). This particular line tells us that in our search for the true $(\beta_{52}^b, \beta_{52}^w)$ point in (b), we can eliminate with certainty all area in the unit square except that on the line, which is clearly an advance over not having the data. Translated into the quantities of interest, this line tells us (by projecting it downward to the horizontal axis) that wherever the true point falls on the line, β_{52}^b must fall in the relatively narrow bounds of [0.07, 0.21]. Unfortunately, in this case, β_i^w can only be bounded (by projecting to the left) to somewhere within the entire unit interval. More generally, lines that are relatively steep, like this one, tell us a great deal about β_i^b and little about β_i^w. Tomography lines that are relatively flat give narrow bounds on β_i^w and wide bounds on β_i^b. Lines that cut off the bottom left (or top right) of the figure give narrow bounds on both quantities of interest.

If the only information available to learn about the unknowns in precinct i is X_i and T_i, a tomography line like that in Figure 0.1 exhausts all this available information. This line immediately tells us the known bounds on each of the parameters, along with the precise relationship between the two unknowns, but it is not sufficient to narrow in on the right answer any further. Fortunately, additional information exists in the other observations in the same data set (X_j and T_j for all $i \neq j$), which, under the right assumptions, can be used to learn more about β_i^b and β_i^w in our precinct of interest.

In order to borrow statistical strength from all the precincts to learn about β_i^b and β_i^w in precinct i, some assumptions are necessary. The simplest version (i.e., the one most useful for expository purposes) of King's model requires three assumptions, each of which can be relaxed in different ways.

First, the set of (β_i^b, β_i^w) points must fall in a single cluster within the unit square. The cluster can fall anywhere within the square; it can be widely or narrowly dispersed or highly variable in one unknown and narrow in the other; and the two unknowns can be positively, negatively, or not at all correlated over i. An example that would violate this assumption would be two or more distinct clusters of (β_i^b, β_i^w) points, as might result from subsets of observations with fundamentally different data generation processes (such as from markedly different regions). The specific mathematical version of this one-cluster assumption is that β_i^b and β_i^w follow a truncated bivariate normal density

$$\mathrm{TN}(\beta_i^b, \beta_i^w | \breve{\mathfrak{B}}, \breve{\Sigma}) = \mathrm{N}(\beta_i^b, \beta_i^w | \breve{\mathfrak{B}}, \breve{\Sigma}) \frac{\mathbf{1}(\beta_i^b, \beta_i^w)}{R(\breve{\mathfrak{B}}, \breve{\Sigma})}, \tag{0.6}$$

where the kernel is the untruncated bivariate normal,

$$\mathrm{N}(\beta_i^b, \beta_i^w | \breve{\mathfrak{B}}, \breve{\Sigma}) = (2\pi)^{-1} |\breve{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2} (\beta_i - \breve{\mathfrak{B}})' \breve{\Sigma}^{-1} (\beta_i - \breve{\mathfrak{B}}) \right], \tag{0.7}$$

and $\mathbf{1}(\beta_i^b, \beta_i^w)$ is an indicator function that equals one if $\beta_i^b \in [0, 1]$ and $\beta_i^w \in [0, 1]$ and zero otherwise. The normalization factor in the denominator, $R(\breve{\mathfrak{B}}, \breve{\Sigma})$, is the volume under

the untruncated normal distribution above the unit square:

$$R(\check{\mathfrak{B}}, \check{\Sigma}) = \int_0^1 \int_0^1 \mathrm{N}(\beta^b, \beta^w | \check{\mathfrak{B}}, \check{\Sigma})\, d\beta^b d\beta^w. \tag{0.8}$$

When divided into the untruncated normal, this factor keeps the volume under the truncated distribution equal to one. The parameters of the truncated density, which we summarize as

$$\check{\psi} = \{\check{\mathfrak{B}}^b, \check{\mathfrak{B}}^w, \check{\sigma}_b, \check{\sigma}_w, \check{\rho}\} = \{\check{\mathfrak{B}}, \check{\Sigma}\}, \tag{0.9}$$

are on the scale of the untruncated normal (and so, for example, $\check{\mathfrak{B}}^b$ and $\check{\mathfrak{B}}^w$ need not be constrained to the unit interval even though β_i^b and β_i^w are constrained by this density).

The second assumption, which is necessary to form the likelihood function, is the absence of spatial autocorrelation: conditional on X_i, T_i and T_j are mean-independent. Violations of this assumption in empirically reasonable (and even some unreasonable) ways do not seem to induce much bias.

The final, and by far the most critical, assumption is that X_i is independent of β_i^b and β_i^w. The three assumptions together produce what has come to be known as the *basic* EI model.[4] King also generalizes this assumption, in what has come to be known as the *extended* EI model, by allowing the truncated normal parameters to vary as functions of measured covariates, Z_i^b and Z_i^w, giving

$$\begin{aligned} \check{\mathfrak{B}}_i^b &= \left[\phi_1(\check{\sigma}_b^2 + 0.25) + 0.5\right] + (Z_i^b - \bar{Z}^b)\alpha^b, \\ \check{\mathfrak{B}}_i^w &= \left[\phi_2(\check{\sigma}_w^2 + 0.25) + 0.5\right] + (Z_i^w - \bar{Z}^w)\alpha^w, \end{aligned} \tag{0.10}$$

where α^b and α^w are parameter vectors to be estimated along with the original model parameters and that have as many elements as Z_i^b and Z_i^w have columns. This relaxes the mean independence assumptions to

$$\begin{aligned} \mathrm{E}(\beta_i^b | X_i, Z_i) &= \mathrm{E}(\beta_i^b | Z_i), \\ \mathrm{E}(\beta_i^w | X_i, Z_i) &= \mathrm{E}(\beta_i^w | Z_i). \end{aligned}$$

Note that this extended model also relaxes the assumptions of truncated bivariate normality, since there is now a separate density being assumed for each observation. Because the bounds, which differ in width and information content for each i, generally provide substantial information, even X_i can be used as a covariate in Z_i. (The recommended default setting in EI includes X_i as a covariate with a prior on its coefficient.) In contrast, under Goodman's regression, which does not include information in the bounds, including X_i leads to an unidentified model (King, 1997: Section 3.2).

These three assumptions – one cluster, no spatial autocorrelation, and mean independence between the regressor and the unknowns conditional on X_i and Z_i – enable one to compute a posterior (or sampling) distribution of the two unknowns in each precinct. A fundamentally important component of EI is that the quantities of interest are not the parameters of the likelihood, but instead come from conditioning on T_i and producing a posterior for β_i^b and β_i^w in each precinct. Failing to condition on T_i and examining the parameters of the truncated bivariate normal only makes sense if the model holds exactly and so is much more

[4] The use of EI to name this method comes from the name of his software, available at `http://GKing.Harvard.edu`.

model-dependent than King's approach. Since the most important problem in ecological inference modeling is precisely model misspecification, failing to condition on T assumes away the problem without justification. This point is widely regarded as a critical step in applying the EI model (Adolph and King, with Herron and Shotts, 2003).

When bounds are narrow, EI model assumptions do not matter much. But for precincts with wide bounds on a quantity of interest, inferences can become model-dependent. This is especially the case in ecological inference problems, precisely because of the loss of information due to aggregation. In fact, this loss of information can be expressed by noting that the joint distribution of β_i^b and β_i^w cannot be fully identified from the data without some untestable assumptions. To be precise, distributions with positive mass over *any* curve or combination of curves that connects the bottom left point ($\beta_i^b = 0,\ \beta_i^w = 0$) to the top right point ($\beta_i^b = 1,\ \beta_i^w = 1$) of a tomography plot cannot be rejected by the data (King, 1997: 191). Other features of the distribution are estimable. This fundamental indeterminacy is of course a problem, because it prevents pinning down the quantities of interest with certainty; but it can also be something of an opportunity, because different distributional assumptions can lead to the same estimates, especially in that only those pieces of the distributions above the tomography lines are used in the final analysis.

0.1.4 King, Rosen, and Tanner's Hierarchical Model

In the continuing search for more information to bring to bear on ecological inferences, King, Rosen, and Tanner (1999) extend King's (1997) model another step. They incorporate King's main advance of combining deterministic and statistical information, but begin modeling a step earlier, at the individuals who make up the counts. They also build a hierarchical Bayesian model, using easily generalizable Markov chain Monte Carlo (MCMC) technology (Tanner, 1996).

To define the model formally, let T_i' denote the *number* of voting age people who turn out to vote. At the top level of the hierarchy they assume that T_i' follows a binomial distribution with probability equal to $\theta_i = X_i\beta_i^b + (1 - X_i)\beta_i^w$ and count N_i. Note that at this level it is assumed that the *expectation* of T_i', rather than T_i' itself, is equal to $X_i\beta_i^b + (1 - X_i)\beta_i^w$. In other words, King (1997) models T_i as a continuous proportion, whereas King, Rosen, and Tanner (1996) recognize the inherently discrete nature of the counts of voters that go into computing this proportion. The two models are connected, of course, since T_i'/N_i approaches θ_i as N_i gets large.

The connection with King's tomography line can be seen in the contribution of the data from precinct i to the likelihood, which is

$$\left(X_i\beta_i^b + (1 - X_i)\beta_i^w\right)^{T_i'} \left(1 - X_i\beta_i^b - (1 - X_i)\beta_i^w\right)^{N_i - T_i'}. \tag{0.11}$$

By taking the logarithm of this contribution to the likelihood and differentiating with respect to β_i^b and β_i^w, King, Rosen, and Tanner show that the maximum of Equation 0.11 is not a unique point, but rather a line whose equation is given by the tomography line in Equation 0.5. Thus, the log likelihood for precinct i looks like two playing cards leaning against each other. As long as T_i is fixed and bounded away from 0.5 (and X_i is a fixed known value between 0 and 1), the derivative at this point is seen to increase with N_i, i.e., the pitch of the playing cards increases with the sample size. In other words, for large N_i, the log likelihood for precinct i degenerates from a surface defined over the unit square into a single playing card standing perpendicular to the unit square and oriented along the corresponding tomography line.

At the second level of the hierarchical model, β_i^b is distributed as a beta density with parameters c_b and d_b, and β_i^w follows an independent beta with parameters c_w and d_w. While β_i^b and β_i^w are assumed *a priori* independent, they are *a posteriori* dependent. At the third and final level of the hierarchical model, the unknown parameters c_b, d_b, c_w, and d_w follow an exponential distribution with a large mean.

A key advantage of this model is that it generalizes immediately to arbitrarily large $R \times C$ tables. This approach was pursued by Rosen, Jiang, King, and Tanner (2001), who also provided a much faster method-of-moments-based estimator. For an application, see King, Rosen, Tanner, and Wagner (2003).

0.2 NEW SOURCES OF INFORMATION IN ECOLOGICAL INFERENCE

We did not attempt to impose an *ex ante* structure on the authors as they were writing, and do not pretend that all the chapters fit into neatly delineated categories. This book is only intended to be a snapshot of a fast-growing field. If you are looking for a textbook, check back in a few years when we have learned more!

Nevertheless, we did need to order the chapters in some way. Our choice was to sort them according to the new sources of information they bring to bear on the ecological inference problem. Thus, Part One offers some alternative baselines that help indicate how much information is lost due to aggregation, and precisely what information is left. For example, in Chapter 1, Jon Wakefield offers a "baseline model," which attempts to make minimal assumptions about individuals and then aggregate up. Remarkably, the likelihood for this model is not flat over the tomography line, even without priors. Similarly, in Chapter 2, Steel, Beh, and Chambers provide a means of formally quantifying the information lost in the aggregation process and thus precisely how much information is left in the aggregate data. They do this through parametric models and hypothesis tests, such as a test for the homogeneity of β_i^b and β_i^w across tables. The authors also show how the increase of even a small amount of information in a standard ecological inference model can greatly improve inferences, even if survey respondents cannot be grouped into precincts or relevant geographic areas. They illustrate their ideas with data from the 1996 Australian census. And finally, in Chapter 3, Stephen Voss shows how the most commonly used method, King's ecological inference model, provides a baseline for understanding and parsing out contextual and compositional effects intertwined in aggregate data.

Part Two of this book is devoted to including sources of information through new models and methods. In Chapter 4, Jeff Lewis finds information where no one had looked before, by including two or more parallel and correlated ecological inference models in the same analysis. His approach, which can be thought of as analogous to a Bayesian version of a "seemingly unrelated regression model," extends King's model by incorporating a key feature of numerous data sources.

In Chapter 5, Bernard Grofman and Samuel Merrill propose three relatively simple methods for ecological inference where the data consist of 2×2 tables. All three introduce new assumptions justified by the authors in terms similar to local smoothing algorithms. The idea is that precincts similar to other precincts on the basis of observables are likely to be similar on unobservables too. The argument introduces a form of information that the authors use to identify where on the tomography lines the point estimates probably lie. The first method is based on minimizing the squared distances from the overall tomography line to each of the precinct-level tomography lines. The other two methods are constrained variants of Goodman regression. Specifically, the second method uses analogous distances to those

used in the first method, but on transformed coordinates rather than the (β_i^b, β_i^w) coordinates. The last method combines Goodman regression with the Duncan–Davis method of bounds. The proposed methods are shown to give answers similar to King's model in several real data sets.

Kevin Corder and Christina Wolbrecht, in Chapter 6, are concerned with estimating newly enfranchised women's turnout in the 1920 U.S. elections in three states. They use the hierarchical Bayesian binomial–normal model proposed by Wakefield but employ informative priors based on prior elections and census data. Their central contribution is to recognize new forms of information in terms of detailed prior, nonsample knowledge of the problem. For example, we know almost for certain that in this period, when women had just gotten the vote, they cast their ballots less frequently than men. In statistical terms, we are essentially certain that $\beta_i^b > \beta_i^w$ for all i and so we can sample from only the portion of the tomography line satisfying the constraint. This greatly increases the information content in their analyses.

In Chapter 7, George Judge, Douglas Miller, and Wendy Tam Cho model the ecological inference problem as an ill-posed inverse problem with a solution selected from the set of feasible solutions – either via maximizing entropy, which implies one set of assumptions, or using the Cressie–Read statistic, which allows for the choice among a variety of others. This approach enables the authors to bring new information to the ecological inference problem in the form of assumptions about individual behavior, often learned from prior survey and other work. The model can be fitted to $R \times C$ tables and allows for explanatory variables reflecting individual spatial or temporal heterogeneity.

In Chapter 8, Ben Pelzer, Rob Eisinga, and Philip Hans Franses propose a model for estimating individual-level binary transitions based on repeated cross-sectional data. The basic problem is equivalent to the classic ecological inference problem with 2×2 tables, where the unknown transition probabilities play the role of the unknown cell probabilities. They introduce assumptions in order to model important information available as lags of some exogenous variables. Inference is performed via maximum likelihood, parametric bootstrap and MCMC methods. The methodology is illustrated with data on personal computer ownership in Dutch households.

Part Three is devoted to methods that attempt to include geographic or time series information in models of ecological inference. In Chapter 10, Kevin Quinn develops Bayesian hierarchical models for ecological inference in the presence of temporal dependence. He builds on Wakefield's approximation to a convolution of binomials and puts priors on the approximate likelihood's parameters reflecting temporal dependence. This class of models may also be useful in some situations for spatial or simultaneous spatiotemporal dependence. Inference is performed via MCMC methods. Quinn studies the methodology via simulated data, as well as by analyzing real data on voting registration by race in Louisiana counties over a 14-year period. Carol Gotway Crawford and Linda Young, in Chapter 10, give an overview of the ecological inference problem from a spatial statistics perspective. These authors point out that ecological inference is a special case of the change-of-support problem in geostatistics, which refers to the geometric size, shape, and spatial orientation of the regions associated with the observed measurements. Changing the support of a variable thus creates a new variable. The problem of how the spatial variation in one variable relates to that in the other is the change-of-support problem, a possible solution being spatial smoothing. The authors illustrate these issues with a case study on low-birth-weight babies.

In Chapter 11, Ernesto Calvo and Marcelo Escolar consider ecological inference in the presence of spatial heterogeneity, which may lead to underestimated standard errors or new forms of bias on top of the aggregation bias inherent in ecological inference. In this chapter

the authors allow for spatial heterogeneity by using geographically weighted regression in the context of Goodman's and King's ecological inference models. Their idea is to incorporate a nonparametric term reflecting spatial effects into these models, resulting in semiparametric models. These models are explored via simulation and with Peronist voting data.

Chapter 12 introduces methods of ecological inference that draw on the extensive spatial epidemiology literature. Therein Sebastien Haneuse and Jon Wakefield show how to model spatial and nonspatial heterogeneity. They incorporate important new information into ecological inference by modeling the fact that multiple diseases share common risk factors, and these risk factors often exhibit spatial clustering. Modeling this clustering, they show, can greatly improve ecological inferences.

Finally, in Part Four, we include comparisons of some existing ecological inference methods. Ruth Salway and Jon Wakefield contrast ecological inference in political science, which tends to focus on descriptive quantities such as the fraction of African Americans voting for the Democrats, and in epidemiology, in which interest is primarily in causal inferences (Chapter 13). Of course, political scientists and most others are also interested in causal inferences, and so the work here should be of general interest. The key problem in making causal inference is confounding, and so Salway and Wakefield analyze the combined effects of confounding bias along with aggregation bias. They show how sources of information about confounding can help improve ecological inferences.

Kenneth Benoit, Michael Laver, and Daniela Giannetti in Chapter 14 discuss the use of King's model in the context of an extensive split-ticket voting application. In Chapter 15, Rogério Silva de Mattos and Alvaro Veiga compare Goodman's regression, King's model, and the hierarchical beta-binomial model (King, Rosen, and Tanner, 1999). To facilitate the simulation-based comparison, the authors use their own version of the beta–binomial model where estimation is performed via the ECM algorithm. The authors' main conclusion is that King's model is superior to the other methods in predictive ability.

In Chapter 16, Micah Altman, Jeff Gill, and Michael McDonald compare the numerical properties of implementations of Goodman regression, King's model, and McCue's method. They look at sources of numerical inaccuracy such as floating point arithmetic, nonlinear optimization, and pseudorandom numbers. The stability and accuracy of the algorithms are tested by introducing random perturbations into the data. The authors' recommendation is to use data perturbations as a diagnostic test in addition to any other diagnostic tools associated with these ecological inference methods.

REFERENCES

Achen, Christopher H. and W. P. Phillips Shively. 1995. *Cross-Level Inference*. Chicago: University of Chicago Press.

Adolph, Christopher and Gary King, with Michael C. Herron and Kenneth W. Shotts. 2003. "A Consensus Position on Second Stage Ecological Inference Models," *Political Analysis*, 11: 86–94.

Duncan, Otis Dudley and Beverly Davis. 1953. "An Alternative to Ecological Correlation," *American Sociological Review*, 18: 665–666.

Goodman, Leo. 1953. "Ecological Regressions and the Behavior of Individuals," *American Sociological Review*, 18: 663–666.

Goodman, Leo. 1959. "Some Alternatives to Ecological Correlation," *American Journal of Sociology*, 64: 610–624.

King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton: Princeton University Press.

King, Gary, Ori Rosen, and Martin A. Tanner, 1999. "Binomial–Beta Hierarchical Models for Ecological Inference," *Sociological Methods and Research*, 28: 61–90.

King, Gary, Ori Rosen, Martin A. Tanner, and Alexander Wagner. 2003. "The Ordinary Election of Adolf Hitler: A Modern Voting Behavior Approach," `http://gking.harvard.edu/files/abs/making-abs.shtml`.

Ogburn, William F. and Inez Goltra. 1919. "How Women Vote: A Study of an Election in Portland, Oregon," *Political Science Quarterly*, 3, XXXIV: 413–433.

Rosen, Ori, Wenxin Jiang, Gary King, and Martin A. Tanner. 2001. "Bayesian and Frequentist Inference for Ecological Inference: The $R \times C$ Case," *Statistica Neerlandica*, 55, 2: 134–156.

Tanner, M. A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd ed., New York: Springer-Verlag.