CROSS-LEVEL/ECOLOGICAL INFERENCE

Wendy K. Tam Cho
Department of Political Science
Northwestern University

Charles F. Manski
Department of Economics and Institute for Policy Research
Northwestern University

The cross-level or ecological inference problem has fascinated scholars for nearly a century (Ogburn and Goltra 1919, Allport 1924, Gehlke and Biehl 1934). The problem occurs when one is interested in the behavior of individuals, but has data only at an aggregated level (e.g., precincts, hospital wards, counties). This data limitation creates a situation where the behavior of individuals must be surmised from data on aggregated sets of individuals rather than on individuals themselves. Since the goal is to make inferences from aggregate units that are often derived from an "environmental level" (i.e. geographical/ecological units such as a county or precinct), the term "ecological inference" is used to describe this type of analysis. Relatedly, while it is often the case that one is interested in individual-level behavior, this problem occurs more generally whenever the level of interest is less aggregated than the level of the data. For instance, one might be interested in behavior at the county level when only state-level data is available. Accordingly, the term "cross-level inference" is often used as a synonym for ecological inference.

The ecological inference problem is an especially intriguing puzzle because it is a very long-standing problem with an exceptionally wide-ranging impact. Occurrences are common across many disciplines, and scholars with diverse backgrounds and interests have a stake in approaches to this problem. Political scientists, for instance, confront these issues when they try to detect whether members of different racial groups cast their ballots differently, using only data at the precinct level that identify vote totals and racial demographics but not vote totals broken down by racial categories. In a completely different substantive area, epidemiologists confront identical methodological issues when they seek to explain which environmental factors influence disease susceptibility using only data from counties or hospital wards, rather than individual patients. Economists studying consumer demand and marketing strategies might need to infer individual spending habits from an analysis of sales data from a specific region and the aggregate characteristics of individuals in that region, rather than from data on individuals' characteristics and purchases. These examples are but a few of the myriad applications and fields where the ecological inference problem has emerged.

Not only does the general cross-level inference problems span many fields, the mathematics of ecological inference are also related to important inferential problems in other disciplines, even when the subject matter is not substantially related. For instance, geographers have long been intrigued with the "modifiable areal unit problem" (MAUP), a problem that is isomorphic to the ecological inference problem. MAUP occurs when the estimates at one level of aggregation are different from the estimates obtained at a different level of aggregation (Yule and Kendall 1950, Openshaw and Taylor 1979). Many statisticians and mathematicians have been captivated by Simpson's Paradox (Simpson 1951), which is the reversal in direction of association between two variables when a third ("lurking") variable is controlled. Described in this way, we can see that Simpson's Paradox (and consequently ecological inference) is akin to the omitted variable problem discussed in virtually all econometrics and regression texts. Scholars with a wide variety of methodological backgrounds and training have simultaneously been contributing to a deeper understanding of the nuances behind making cross-level inferences. Their notation and terminology may differ, but the similarity of the underlying problem cannot be denied. These connections are important to note, since scholars so often gain their keenest insight into how to approach a problem of interest by making a foray into an established literature in a field afar from their own. The value of tying together methodological developments across disciplinary boundaries can be enormous.

This chapter is intended to be an exposition of some of the main methodological approaches to the ecological inference problem. We present our discussion in two parts. We first pass through a general exposition, with minimal math, and then approach the problem with significantly more mathematical detail. In the first part, we begin by discussing the fundamental indeterminacy of the problem. We then present a framework that coherently binds the variety of approaches that have been proposed to address this problem. Next, we provide an overview of these various approaches and comment on their respective contributions. In the second part, we place the ecological inference problem within the literature of partial identification and discuss recent work generalizing the use of logical bounds on possible solutions as an identification region for the general $r \times c$ problem. Finally, we conclude, cautiously but optimistically, with some admonitions about this fascinating problem that has enthralled decades of scholars from varied disciplines.

1 Ecological Inference Then and Now

1.1 Fundamental Indeterminacy

The ecological inference problem is an example of an inverse problem with many "solutions." Accordingly, the problem is defined by ("plagued with" one might lament) a fundamental indeterminacy. Although one may wish to obtain a point estimate or "solution" to the problem, there is not enough information to narrow down the feasible set of estimates to one particular point estimate. This type of problem has been described as "partially identified" because while the available information does not allow one to identify the parameter of interest without the imposition of strong assumptions, it does often allow one to identify a "region" within which the parameter must lie, thus partially identifying the problem. How to proceed with partially identified problems such as the ecological inference problem is not obvious.

For consistency, ease of illustration, and without loss of generality, we will employ the example of racial voting in the United States throughout this chapter. Parallels to other applications should not be difficult to spot. In this voting application, the challenge is to infer how individuals voted from election returns aggregated at the precinct level. Because U.S. elections employ the secret ballot, individual vote choices are unknown. Election returns, however, are reported at the precinct level, and aggregate individual racial categorizations can usually be obtained and merged with the voting data. Hence, for any given precinct, the available data include how many votes each candidate received as well as the precinct's racial composition. The problem can be summed up with a contingency table akin to the one shown in Table 1.[1] Here, Y represents vote choice, Z represents racial group, and X represents the precinct. This contingency table represents a single particular precinct. There are I similar tables to represent each of I total precincts in a district.

In this example, there are 2 racial groups and 3 candidates, presenting a 2×3 problem. Y takes one of 3 values, y_1, y_2, or y_3; Z takes one of 2 values, z_1 or z_2; and X takes one of I values, $x_1, \ldots, x_I$. The known elements of the table are displayed as the so-called marginals, the proportion of the vote each candidate received in a precinct, $\mathrm{P}(Y \mid X)$, and the proportion of each racial group in a precinct, $\mathrm{P}(Z \mid X)$. The interior cells are conditional probabilities, that is, the probability that we observe a vote choice given a precinct and racial group, $\mathrm{P}(Y \mid Z, X)$. Political scientists may be more familiar with notation resembling the lower contingency table in Table 1. Here, the βs are often referred to as parameters, but they are equivalent to the conditional probabilities shown in the upper contingency table. That is, $\beta_{ij} = \mathrm{P}(Y = y_j \mid Z = z_i, X)$. We know, by Bayes's Theorem, that $\beta = \mathrm{P}(Y, Z \mid X) = \mathrm{P}(Y \mid Z, X)\,\mathrm{P}(Z \mid X)$. Since $\mathrm{P}(Z \mid X)$, the proportion of voters in each racial group, is observed via census data, the ecological inference problem boils down to inferring $\mathrm{P}(Y \mid Z, X)$, which is not observed. If the ballot had one fewer candidate, it would pose a 2×2 problem, which would be appealing because the mathematics involved for a 2×2 problem are simpler than those for an $r \times c$ problem, where $r > 2$ and/or $c > 2$. However, whether the problem is 2×2 or $r \times c$, the basic inferential problem, which is to infer the conditional probabilities, $\mathrm{P}(Y \mid Z, X)$, from the

[1] Note here that we assume that everyone turned out to vote. Of course, this is an unrealistic assumption. In real American elections, substantial proportions of each voter group abstain from voting, which means there is one additional category of vote and the ecological inference problem is more complicated computationally, though not conceptually. For present purposes, ignoring abstention reduces clutter and obscures nothing about the nature of the inference problem. Note also that there is a literature discussing this problem of measurement error (i.e. assuming that our measure of turnout is perfect) in ecological election analysis (Kousser 1973, Loewen 1982, Kleppner 1985, and Grofman, Migalski, and Noviello 1985).

Table 1: **Contingency Table for precinct** $X = x_1$

	Candidate 1	**Candidate 2**	**Candidate 3**	
Group 1	$\mathrm{P}(Y = y_1 \mid Z = z_1, X)\,\mathrm{P}(Z = z_1 \mid X)$	$\mathrm{P}(Y = y_2 \mid Z = z_1, X)\,\mathrm{P}(Z = z_1 \mid X)$	$\mathrm{P}(Y = y_3 \mid Z = z_1, X)\,\mathrm{P}(Z = z_1 \mid X)$	$\mathrm{P}(Z = z_1 \mid X)$
Group 2	$\mathrm{P}(Y = y_1 \mid Z = z_2, X)\,\mathrm{P}(Z = z_2 \mid X)$	$\mathrm{P}(Y = y_2 \mid Z = z_2, X)\,\mathrm{P}(Z = z_2 \mid X)$	$\mathrm{P}(Y = y_3 \mid Z = z_2, X)\,\mathrm{P}(Z = z_2 \mid X)$	$\mathrm{P}(Z = z_2 \mid X)$
	$\mathrm{P}(Y = y_1 \mid X)$	$\mathrm{P}(Y = y_2 \mid X)$	$\mathrm{P}(Y = y_3 \mid X)$	

	Candidate 1	**Candidate 2**	**Candidate 3**	
Group 1	$\beta_{11}z_1$	$\beta_{12}z_1$	$\beta_{13}z_1$	z_1
Group 2	$\beta_{21}z_2$	$\beta_{22}z_2$	$\beta_{23}z_2$	z_2
	y_1	y_2	y_3	

Table 2: **2 × 2 Contingency Table**

	Candidate 1	**Candidate 2**	
Group 1	$\beta_{11}z_1$	$(1-\beta_{11})z_1$	z_1
Group 2	$\beta_{21}z_2$	$(1-\beta_{21})z_2$	z_2
	y_1	y_2	

observed values, $\mathrm{P}(Y \mid X)$ and $\mathrm{P}(Z \mid X)$, does not change. The core issue is that multiple values of the former probabilities are consistent with given values for the latter ones.

We use the simple 2×2 example shown in Table 2 to illustrate the indeterminacy. The known elements of this problem are the proportion of voters in each group, z_1 and z_2, and the proportion of votes received by each candidate, y_1 and y_2. The unknown elements are the proportions of each group that voted for each candidate, β_{11}, β_{12}, β_{21}, and β_{22}. If we assume that group 1 and group 2 are mutually exclusive and exhaustive, then z_2 is simply $(1 - z_1)$. Similarly, if we assume that all votes were cast for either candidate 1 or candidate 2 and there were no abstentions, then $y_2 = (1 - y_1)$. Usually, there is such a contingency table to describe each areal unit in the data set. In our voting application, since a set of precincts comprise the voting district in question, one might want to index these values with a further subscript to indicate the precinct. For each precinct, i, then, we have the following relationship,

$$y_{1i} = \beta_{11i}z_{1i} + \beta_{21i}(1 - z_{1i})$$

This relationship holds exactly for each of the I precincts in the district, yielding a system with I equations (one for each precinct) and $2I$ unknowns (two parameters, β_{11i} and β_{21i}, for each precinct). The available information is that these $2I$ parameters solve the I equations and, moreover, that each parameter takes a value between 0 and 1. The fundamental indeterminacy is that multiple values of the parameters satisfy these restrictions. Note as well that adding observations of further precincts does not change the indeterminacy. Each new precinct adds one more equation to the system and two more parameters to be inferred.

1.2 Various Approaches

Despite the fundamental indeterminacy, scholars have nonetheless pursued the ecological inference problem. Certainly, the substantive applications are interesting and important. Although the multifarious approaches to this problem have been distinct, they can be seen as lying along a continuum. The left end of the continuum is characterized by a lack of assumptions, and, concomitantly, high credibility. Even when

Table 3: **Duncan and Davis Bounds Example: Employed Females by Occupation and Color for Chicago, 1940**

	Not Domestic Service	Domestic Service	
White	[348,578, 376,179] [0.912 , 0.984]	[6,073, 33,674] [0.016 , 0.088]	382,252 (0.933)
Nonwhite	[0, 27,601] [0.00, 1.00]	[0, 27,601] [0.00 , 1.00]	27,601 (0.067)
	376,179 (0.918)	33,674 (0.082)	409,853

no assumptions are imposed, the data can allow one to narrow the range in which the true values of the parameters of interest lie (Duncan and Davis 1953). To move along this continuum, assumptions must be made. At the right end of the continuum are approaches that lead one to a precise point estimate for each parameter. There are, to be sure, many different approaches clustered on the right, in the region of strong and numerous assumptions. In general, one cannot zero in on precise estimates without making restrictive assumptions, and thus trading reduced credibility for increased "precision." Whether or how one settles on a model, with attendant assumptions, in order to narrow the range of estimates is a critical decision that is inextricably tied to the eventual substantive interpretation of the analysis. How to proceed is rarely obvious on the basis of theory alone, and what criteria one should employ is a function of the application at hand, the data, and the costs and benefits associated with different kinds of inferential errors. There are no pat answers, no generic solutions. Instead, each individual project of data analysis merits careful and explicit consideration.

1.2.1 Ranges

In some applications, simply narrowing the range in which the possible estimates lie may be sufficient. Duncan and Davis (1953) developed the idea of bounds in an ecological inference problem. Table 3 reproduces their example from data for Chicago in 1940.[2] The contingency table shows the breakdown of race and occupation for females. Usually, we can observe only the marginals of this table. We need to infer the unobservable values in the interior cells from the observed values in the marginals. Given only the marginals, we can surmise nothing more than that the percentage of nonwhites engaged in domestic service spans the entire range from 0% to 100%. However, with 33,674 women in domestic service total, and

[2] Note that the interior cells of Table 3 are set up a bit differently than Table 1 and 2. The interior cells of Table 3 give the possible range of the overall number of persons in the cell. We present these different, though obviously related, interior cells to maximize similarity with the Duncan and Davis presentation and to simplify the discussion of how to compute ranges.

only 27,601 nonwhite women, there cannot be fewer than 6,073 (33,674 – 27,601) white women in domestic service. Hence, the range of whites engaged in domestic service is fairly narrow (1.6% to 8.8%). Similarly, the range of whites engaged in non-domestic service is small (91.2% to 98.4%). Depending on one's interest with these data, these computed ranges may be sufficient for one's substantive interest in the problem. In the Duncan and Davis framework, the greatest appeal is that no assumptions need be made. Clearly, the best possible outcome is attained when nothing is assumed to define the range of possible estimates, and these computed bounds are sufficient to make the type of inferences that are warranted. Moreover, these bounds are simple to compute:

$$\beta_{11} \in \left[\max\left(0, \frac{Y_1 - (1 - Z_1)}{Z_1}\right), \min\left(\frac{Y_1}{Z_1}, 1\right)\right]$$

and

$$\beta_{21} \in \left[\max\left(0, \frac{Y_1 - Z_1}{1 - Z_1}\right), \min\left(\frac{Y_1}{1 - Z_1}, 1\right)\right].$$

It is important to note that even wide bounds may be helpful. This is particularly true if one is interested in gaining a sense of how informative the data are or how much one might need to rely on assumptions to move toward a point estimate. How "wide" or "narrow" bounds are is dependent upon the available data and what one wishes to discover. Indeed, a point estimate is not always the end goal. Instead, the true goal is to seek insight into the substantive question of interest. A point estimate (with questionable reliability) is one way of obtaining such insight, but there are other vehicles as well, and having a firm grasp of how informative the data are is an important aspect of the analysis. If one moves directly to a single estimator that yields a point estimate, one is wedded to the strong assumptions that were required to identify the problem. Hence, even though it is rare in real applications to obtain tight bounds on all parameters and it would be unusual that the bounds were sufficient to address all of one's substantive interests, the inherent value of the bounds should not be underestimated. It provides a basis from which to understand how informative the data are without the imposition of questionable assumptions and serves as a domain of consensus among all researchers who may otherwise disagree on the appropriate assumptions for an application.

1.2.2 Point Estimates

Although estimating ranges is important and easily integrated in an ecological inference analysis, the information emerging from such an analysis may be insufficiently direct on the substantive question of interest. It may be the case that one needs to narrow the feasible set of estimates further, indeed, even to a single point estimate. If so, one has the option of employing an ecological inference model that will yield a point estimate. The tie that binds all models that yield a point estimate is that they rely heavily on strong as-

Table 4: **Assumptions**

Precinct 1

	Democrat	**Republican**	
Black			150 (0.30)
White			350 (0.70)
	300 (0.60)	200 (0.40)	500

Precinct 2

	Democrat	**Republican**	
Black			350 (0.449)
White			430 (0.551)
	400 (0.513)	380 (0.487)	780

sumptions. In any particular application, one may have a sense for which assumptions are more tenable, but whether those assumptions hold is never known with certainty. The trade off is that one obtains a point estimate, which is helpful, but if the assumptions required to narrow the feasible estimates to a single point are wrong, the point estimate will be incorrect and misleading. The potential downside is significant. Accordingly, employing any ecological inference model that yields a point estimate should be approached guardedly with attention to the impact of the assumptions.

Following Robinson's (1950) seminal article demonstrating the lack of agreement between ecological correlation and individual correlation, and the articles which followed Robinson still using ecological correlations as a substitute for individual correlations, Goodman (1953) sought to clarify the problem in the regression context. He cautioned that "*in general* the study of the regression between ecological variables cannot be used as substitutes for the study of the behavior of individuals" (1953, 663). However, "*in very special circumstances* the study of regression between ecological variables may be used to make inferences concerning the behavior of individuals" (1953, 663). His key point is that for regression to yield insight into individual-level behavior, the "constancy assumption" must hold. That is, in our running example, voters in a particular racial group vote similarly regardless of precinct of residence. Goodman (1953, 1959) did not advocate the use of regression for ecological inference, but instead sought to illuminate the strong assumptions required to make such an exercise fruitful.

Consider the two contingency tables in Table 4 where the left contingency table represents one precinct, call it "Precinct 1," and the right contingency table represents a second precinct ("Precinct 2"). Consider just the left contingency table for the moment. How might one determine the values of the interior cells given only the information in the marginals? If any one of the four interior cells were known, then one could determine the other three. For instance, if we knew that 100 blacks voted for the Democratic candidate, then $(150 - 100) = 50$ blacks must have voted for the Republican candidate. In a similar calculation, we could determine that $(300 - 100) = 200$ whites voted for the Democratic candidate and so $(350 - 200) = 150$

whites voted for the Republican candidate. If we can set any one of the interior cells to a specific value, the other cells can be determined easily. Without any further information than that contained in the marginals, however, it is impossible to determine a specific value for any of the interior cells with certainty. At this juncture, imposing assumptions is a necessity to move toward a point estimate.

Imposing the Goodman "constancy assumption" (voters in a particular racial group vote similarly regardless of precinct of residence) is one of an innumerable number of ways to proceed. Following this approach, if 50%, say, of blacks voted for the Democratic candidate in precinct 1, then outside of random variation, 50% of the blacks in precinct 2 voted for the Democratic candidates. Race is a factor associated with vote choice, but precinct, in effect, is not relevant. With our two example precincts, we have a system of two equations:

$$\begin{aligned} 0.6 &= 0.3\,\beta_{11} + 0.7\,\beta_{21} \\ 0.513 &= 0.449\,\beta_{12} + 0.551\,\beta_{22}, \end{aligned} \tag{1}$$

where β_{1j} is the proportion of blacks voting for the Democratic candidate in precinct j and β_{2j} is the proportion of whites voting for the Democratic candidate in precinct j. As we can see, we have 2 equations and 4 unknowns. If we impose the constancy assumption across precincts, then $\beta_{11} = \beta_{12}$ and $\beta_{21} = \beta_{22}$, and we can create a system with 2 equations and 2 unknowns, $\beta_1 = \beta_{11} = \beta_{12}$ and $\beta_2 = \beta_{21} = \beta_{22}$. It is clear that the system with 2 equations and 2 unknowns,

$$\begin{aligned} 0.6 &= 0.3\,\beta_1 + 0.7\,\beta_2 \\ 0.513 &= 0.449\,\beta_1 + 0.551\,\beta_2 \end{aligned} \tag{2}$$

has a unique solution. Simple algebra leads to the solution, $\beta_1 = 0.1913$ and $\beta_2 = 0.7752$.

While imposing the constancy assumption leads to a "solution," it is questionable in our application, since it implies that contextual factors are irrelevant. The implication is that whites in homogeneously white neighborhoods have exactly the same vote preferences as whites living in predominantly black or otherwise racially diverse neighborhoods. The most appealing aspect of this assumption is that the parameters are identified and a point estimate can be obtained if this assumption is imposed. This appeal is greatly lessened by the least appealing aspect of this identifying assumption, which is that the point estimate may be wrong and grossly misleading.

The role of assumptions is always fundamental, but has been more obvious in some models and less obvious in others. Whether obvious or not, the basic tenets behind the role of assumptions and what they entail for estimation are unyielding. For instance, while the King (1997) ecological inference model is laden with assumptions, the impact and role of these assumptions in determining the point estimate is not transparent. Cho (1998) demonstrates that when the strong assumptions of the King model hold, then that model is appropriate. When the assumptions of the King model are violated, the resulting point estimates and their associated measures of uncertainty are unreliable (see also Rivers 1998, Freedman et al. 1998, 1999, McCue 2001, and Cho and Gaines 2004). A key assumption behind the King model is the distributional assumption. King does not impose a "constancy assumption," but his distributional assumption that the parameters or conditional probabilities follow a truncated bivariate normal distribution (TBN) is akin to a "similarity assumption." That is, the percentage of blacks who vote for the Democrat need not be the same across precincts, but the set of estimates for Democratic voting by blacks should follow a truncated bivariate normal distribution so that in most precincts, the percentage congregates around the mean of the TBN. King does not offer any empirical evidence that this distributional assumption is tenable, but he does posit that "[s]ince most areal units were created for administrative convenience and aggregate individuals with roughly similar patterns, ecological observations from the same data sets usually do have a lot in common. Even though Goodman's assumption that [the parameters] are constant over i is usually false, the assumption that they vary but have a single mode usually fits aggregate data" (King 1997, 191–192). Interestingly, Monte Carlo simulations indicate that the King model and OLS yield virtually identical estimates in the vast majority of cases (Cho and Yoon 2001, Anselin and Cho 2002), implying that for substantive purposes, the "similarity assumption" does not differ wildly from the constancy assumption.

Freedman et al. (1991) proposed the neighborhood model to demonstrate the role of assumptions in ecological inference models. His model assumes that voters within a precinct, regardless of race, vote similarly. They did not propose this model as a serious attempt to extract individual-level behavior from aggregate data, but rather to demonstrate that one can arrive at point estimates in a variety of ways and the assumptions that one incorporates in this process heavily influence the eventual outcome. The neighborhood model, in effect, imposes a different type of constancy assumption. Instead of constancy across precincts, it imposes constancy across racial groups. In our simple two-precinct example, then, $\beta_1 = \beta_{11} = \beta_{21}$ and $\beta_2 = \beta_{12} = \beta_{22}$. This approach is another way to convert our system of 2 equations and 4 unknowns into a system with 2 equations and 2 unknowns. The system (2) then becomes the following

$$\begin{aligned} 0.6 &= 0.3\,\beta_1 + 0.7\,\beta_1 \\ 0.513 &= 0.449\,\beta_2 + 0.551\,\beta_2 \end{aligned} \tag{3}$$

and the solution is then $\beta_1 = 0.6$ and $\beta_2 = 0.513$. It is simple to see how we can arrive at a unique solution by imposing some type of assumption or constraint. Indeed, there are a very large number of different assumptions that can be imposed and that the assumptions completely determine the "solution."

Table 5: **Assumptions**

	Democrat	**Republican**	
Black	30	120	150
White	270	80	350
	300	200	500

	Democrat	**Republican**	
Black	70	85	150
White	230	120	350
	300	200	500

To be sure, many have ventured to explore and propose various assumptions and models for ecological inference. For instance, Hawkes (1969) proposed an aggregated multinomial model where the parameters have fixed multinomial distributions over some number of precincts. Brown and Payne (1986) introduced a random effects model with a specified covariance structure, allowing the parameters to vary over precincts through dependence on precinct-specific covariates and modeling any additional variance through their aggregated compound multinomial distribution. Crewe and Payne (1976) suggested a multiple regression model to account for remaining variation after controlling for race. Calvo and Escolar (2003) presented geographically weighted regressions as a way to account for spatial variation. Indeed, the role of spatial dependence has attracted a flury of recent scholarship (see, e.g., Anselin and Cho 2002, Wakefield 2003, 2004). Imai and Lu (2005) suggest formulating the problem as a missing data problem. Thomsen (1987) advanced a latent variable logit model where the latent variable is normally distributed and the logits of the marginal values depend linearly on the latent variable. As in some of the previously mentioned models, the Thomsen model assumes that there are some subsets of homogeneous precincts. This is a short list and not nearly an exhaustive list. Since there are innumerable assumptions that one could impose, the possibilities are limitless.

Cho, Judge, and Gaines (2005) highlight the philosophy behind model building in ill-posed situations such as ecological inference. They contend that the most logical way to think through this modeling challenge is to begin with a minimum number of assumptions. The principle is that one should not assume more than one knows, since assumptions, after all, may be incorrect. In our example precinct 1 in Table 4, it is obvious that many different values will be consistent with the observed marginal values. For instance, two possibilities are shown in Table 5. How many ways are there to partition the 150 black voters into 30 Democratic votes and 120 Republican votes? This is a simple combinatorial problem, and the answer is $\binom{150}{30}$. Similarly, there are $\binom{350}{270}$ ways to partition the 350 white voters such that there are 270 Democratic votes and 80 Republican votes. So, the number of ways in which we might observe the outcome shown in the left-most table is $\binom{150}{30} \times \binom{350}{270} \approx 8.3 \times 10^{111}$. Similar calculations show that the right-most table is an outcome that could occur with $\binom{150}{70} \times \binom{350}{230} \approx 1.6 \times 10^{140}$ different partitioning of the available voters.

Since this latter number is greater than the former number, if we impose no other assumptions, then the outcome in the right-most table must be preferred to the outcome in the left-most table because it can occur in a greater number of ways. Clearly, there is a very large number of outcomes that are consistent with the marginals. However, after assuming as little as possible, the more logical point estimate must be the one most strongly indicated by the data, that is, the point estimate that could occur in the greatest number of ways. This "solution" is not a "magic bullet" in that it will always give the right answer, it is simply a rule for inductive reasoning that leads to the point estimate most strongly indicated by the data at hand.

Lastly, another approach, taken by Cho (2001), is to develop tests for the assumptions underlying the models and to adjust appropriately. So, if the constancy assumption appears untenable, perhaps one should test whether the assumption may hold. This might be done in a manner similar to the way that others have tested for parameter constancy in the switching regressions context. The idea is that voters in racially homogeneous precincts may have differ preferences than voters in racially diverse precincts. So, whites in homogeneously white precincts may behave similarly to each other but distinct from whites in racially diverse precincts. In this sense, the constancy assumption may hold among different subsets of data. If one can identify these subsets where constancy can be reasonably assumed, then one can proceed in the usual regression context within homogeneous sets of precincts.

Several works are also helpful in summarizing estimators and discussing the various particularities of their performance. Cleave, Brown, and Payne (1995) and Achen and Shively (1995) are especially noteworthy on this score. These lively debates and discussions have also been visited in other disciplines (see, e.g., Jones 1972, Kousser 1973, and Lichtman 1974) as well as across disciplines (Salway and Wakefield 2004). Hanushek et al. (1974) and Achen and Shively (1995) are also very helpful in illuminating the issues that surround model specifications for the individual and aggregate data and in discussing how the specification for those two levels are related or, unintuitively, unrelated.

2 Ecological Inference as a Problem of Partial Identification

Each of these many varied approaches to the ecological inference problem brings a unique insight into the problem. In this Section, we highlight an under-visited vantage point from which ecological inference can be viewed. In particular, we discuss the ecological inference problem as a specific instance of a problem of *partial identification of probability distributions*, where the sampling process generating the available data does not fully reveal a distribution of interest. Other familiar partial identification problems include inference with missing data or with errors in measurement, as well as inference on treatment response from observational data. These are problems of identification, rather than of statistical inference, because they persist as sample size increases. They are problems created by the type of data that are available, rather than by the quantity of data. There is a recent and fast growing literature on partial identification, insights from which are clearly applicable, but underexplored by political scientists. We now turn to discussing how insights from this literature can and should be infused into the study of the ecological inference problem.

A first principle in studying problems of partial identification might be to determine what the sampling process does reveal about the distribution of interest (Manski 1995, 2003). In ecological inference, this means finding a range for the feasible set of estimates. Using the constraints (in our case, the marginal values), we can determine that the distribution of interest must lie in a specific set of feasible distributions, its *identification region*. As we have discussed in Section 1.2.1, the identification region may be sufficient for our substantive concerns, but it may not be. Depending on the application at hand, one might proceed to explore what more can be learned when the available data are combined with assumptions that are credible enough to be taken seriously. We say that such assumptions have *identifying power* if they narrow the identification region. Sufficiently strong assumptions may *point-identify*, that is fully reveal, the distribution of interest.

Over fifty years ago, the modern literature on ecological inference began to traverse the path suggested by the partial identification literature when Duncan and Davis (1953) demonstrated that the data available in ecological inference imply the bounds on the probabilities $\mathrm{P}(Y \mid X, Z)$ that we stated in Section 1.2.1. Contemporaneously, Goodman (1953) showed that the data combined with a specific constancy assumption can point-identify $\mathrm{P}(Y \mid X, Z)$, as we showed in Section 1.2.2. However, further work along these lines has not been systematically pursued despite an obvious connection to the partial identification literature. Moreover, these pioneering researchers studied only the relatively simple $2{\times}2$ case in which y and z are both binary variables. The general inferential problem has been addressed only recently, in Cross and Manski (2002). We combine their main ideas and findings with our running example to illuminate the manner in which these two literatures are intertwined.

2.1 Inference Using the Data Alone

As we stated earlier, the parameters in an ecological inference application can be understood as conditional probabilities. This framework meshes with the partial identification literature seamlessly, and we continue with the same notation. In particular, let each member j of population **J** have an outcome y_j in a space **Y** and covariates (x_j, z_j) in a space $\mathbf{X} \times \mathbf{Z}$. Let the random variable $(Y, X, Z) : \mathbf{J} \rightarrow \mathbf{Y} \times \mathbf{X} \times \mathbf{Z}$ have distribution $\mathrm{P}(Y, X, Z)$. The general goal is to learn the conditional distributions $\mathrm{P}(Y \mid X, Z) \equiv \mathrm{P}(Y \mid X = x, Z = z), (x, z) \in \mathbf{X} \times \mathbf{Z}$. A particular objective may be to determine the mean regression $\mathrm{E}(Y \mid X, Z) \equiv E(Y \mid X = x, Z = z), (x, z) \in \mathbf{X} \times \mathbf{Z}$. We assume only that X and Z are discrete variables, with $\mathrm{P}(X = x, Z = z) > 0$ for all $(x, z) \in \mathbf{X} \times \mathbf{Z}$. The variable Y may be either discrete or continuous.

The joint realizations of (Y, X, Z) are not observable, but data are available from two sampling processes. One process draws persons at random from **J** and generates observable realizations of (Y, X) but not Z. In our example, voting records reveal (Y, X), the vote received by a certain candidate in a given precinct. The other sampling process draws persons at random and generates observable realizations of (X, Z) but not Y. These data are available from merging racial demographic data from the census with precinct boundaries. The two sampling processes reveal the distributions $\mathrm{P}(Y, X)$ and $\mathrm{P}(X, Z)$. Ecological inference is the use of this empirical evidence to learn about $\mathrm{P}(Y \mid X, Z)$, the propensity to vote for a certain candidate conditional on precinct and race, an unobservable behavior because of the secret ballot. The structure of the ecological inference problem is displayed by the Law of Total Probability

$$\mathrm{P}(Y \mid X) = \sum_{z \in \mathbf{Z}} \mathrm{P}(Y \mid X, Z = z)\ \mathrm{P}(Z = z \mid X). \tag{4}$$

While the bounds, or identification region for the 2×2 case are relatively simple to compute, the problem becomes much more complex in the $r \times c$ case. We summarize Cross and Manski (2002), who have provided the identification region for the general $r \times c$ case, denoted $\mathbf{H}[\mathrm{P}(Y \mid X = x, Z)]$. In particular, let $\mathbf{\Gamma}_{\mathrm{Y}}$ denote the space of all probability distributions on **Y**. Let $x \in \mathbf{X}$. Define $\mathrm{P}(Y \mid X = x, Z) \equiv [\mathrm{P}(Y \mid X = x, Z = z), z \in \mathbf{Z}]$. Let $\mid \mathbf{Z} \mid$ be the cardinality of **Z**. Then a $\mid \mathbf{Z} \mid$-vector of distributions $[\eta_z, z \in \mathbf{Z}] \in (\mathbf{\Gamma}_{\mathrm{Y}})^{|\mathbf{z}|}$ is a feasible value for $\mathrm{P}(Y \mid X = x, Z)$ if and only if it solves the finite mixture problem

$$\mathrm{P}(Y \mid X = x) = \sum_{z \in \mathbf{Z}} \eta_z \mathrm{P}(Z = z \mid X = x). \tag{5}$$

It follows that the identification region for $\mathrm{P}(Y \mid X = x, Z)$ using the data alone is the set

$$\mathbf{H}[\mathrm{P}(Y \mid X = x, Z)] \equiv \{(\eta_z, z \in \mathbf{Z}) \in (\mathbf{\Gamma}_{\mathrm{Y}})^{|\mathbf{z}|} : \mathrm{P}(Y \mid X = x) = \sum_{z \in \mathbf{Z}} \eta_z\ \mathrm{P}(Z = z \mid X = x)\}. \tag{6}$$

Moreover, the identification region for $\mathrm{P}(Y \mid X, Z)$ is the Cartesian product $\times_{x \in \mathbf{X}}\ \mathbf{H}\,[\mathrm{P}(Y \mid X = x, Z)]$. This holds because the Law of Total Probability (4) restricts $\mathrm{P}(Y \mid X, Z)$ across values of Z only, not across values of X. Equation (6) is simple in form but is too abstract to communicate much about the size and shape of the identification region. Hence, practical application requires further study.

A relatively simple result emerges if the objective is inference on $\mathrm{P}(Y \mid X = x, Z = z)$ for one specified covariate value $(x, z) \in \mathbf{X} \times \mathbf{Z}$, or just one precinct and race. In this instance, let $p \equiv \mathrm{P}(Z \neq z \mid X = x)$. Cross and Manski show that the identification region for $\mathrm{P}(Y \mid X = x, Z = z)$ is the set

$$\mathbf{H}\,[\mathrm{P}(Y \mid X = x, Z = z)] = \mathbf{\Gamma}_{\mathrm{Y}}\ \cap\ [\mathrm{P}(Y \mid X = x) - p\,\gamma]/(1 - p), \gamma \in \mathbf{\Gamma}_{\mathrm{Y}}. \tag{7}$$

In the special case where Y is binary, this is equivalent to the Duncan and Davis bounds.

Characterization of the identification region is much more involved when the objective is joint inference on the full vector of conditional distributions $\mathrm{P}(Y \mid X = x, Z)$, or on the full set of races residing in a precinct. Cross and Manski address an important aspect of this question—the identification region for the mean regression $\mathrm{E}(Y \mid X = x, Z)$. Equation (6) implies that the feasible values of $\mathrm{E}(Y \mid X = x, Z)$ are

$$\mathbf{H}\left[\mathrm{E}(Y \mid X = x, Z)\right] = \{\left(\textstyle\int y \, \mathrm{d}\eta_z, z \in \mathbf{Z}\right), (\eta_z, z \in \mathbf{Z}) \in \mathbf{H}\left[\mathrm{P}(Y \mid X = x, Z)\right]\}. \tag{8}$$

Cross and Manski characterize $\mathbf{H}[\mathrm{E}(Y \mid X = x, Z)]$ less abstractly than (8) by demonstrating that this is a convex set having finitely many extreme points, which are the expectations of certain $\mid \mathbf{Z} \mid$-tuples of *stacked distributions*. Their result, although still somewhat complex, provides a constructive way to compute the identification region or the bounds for the general $r \times c$ case.

2.2 Assumptions and Instrumental Variables

As we have earlier discussed, bounds or identification regions using the data alone may be helpful, but there are certainly applications where one wishes to reduce the feasible set of estimates further, and perhaps even to a single point estimate. In these cases, one must proceed to impose assumptions. The Goodman constancy assumption is an application of the broad idea of *instrumental variables*, which has been used widely in econometrics from the 1940s onward. Cross and Manski consider two assumptions that use components of X as instrumental variables. Let $X = (V, W)$ and $\mathbf{X} = \mathbf{V} \times \mathbf{W}$. One could assume that Y is mean-independent of V, conditional on (W, Z); that is,

$$\mathrm{E}(Y \mid X, Z) = \mathrm{E}(Y \mid W, Z). \tag{9}$$

Alternatively, one could assert that Y is statistically independent of V, conditional on (W, Z); that is,

$$\mathrm{P}(Y \mid X, Z) = \mathrm{P}(Y \mid W, Z). \tag{10}$$

Both assumptions use V as an instrumental variable. Assumption (10) is stronger than (9) unless Y is binary, in which case (9) and (10) are equivalent. In the 2×2 case where Y and Z are both binary, (9) is Goodman's constancy assumption.

Let $w \in \mathbf{W}$. The identification regions for $\mathrm{E}(Y \mid W = w, Z)$ under assumptions (9) and (10) can respectively be shown to be

$$\mathbf{H}^{*}_{w} \equiv \bigcap_{v \in \mathbf{V}} \mathbf{H}\left[\mathrm{E}(Y \mid V = v, W = w, Z)\right] \tag{11}$$

and

$$\mathbf{H}_w^{**} = \{\left(\int y \, \mathrm{d}\eta_z, z \in \mathbf{Z}\right), (\eta_z, z \in \mathbf{Z}) \in \bigcap_{v \in \mathbf{V}} \mathbf{H}\left[\mathrm{P}(Y \mid V = v, W = w, Z)\right]\}. \tag{12}$$

The corresponding identification regions for $\mathrm{E}(Y \mid W, Z)$ are $\times_{w \in \mathbf{w}} \mathbf{H}_w^*$ and $\times_{w \in \mathbf{w}} \mathbf{H}_w^{**}$. Observe that set $\mathbf{H}_w^*$ and/or $\mathbf{H}_w^{**}$ could turn out to be empty. If this occurs, we can conclude that assumption (9) and/or (10) must be incorrect.

Equations (11) and (12) are too abstract to convey a sense of the identifying power of assumptions (9) and (10). Cross and Manski show that a simple *outer identification region* (that is, a set containing the identification region) emerges if one exploits only the Law of Iterated Expectations rather than the full force of the Law of Total Probability. Let assumption (9) hold. Let $w \in \mathbf{W}$. Let $\mid \mathbf{V} \mid$ denote the cardinality of $\mathbf{V}$. Let $\pi_{(v,w)z} \equiv \mathrm{P}(Z = z \mid V = v, W = w)$ and let $\mathbf{\Pi}$ denote the $\mid \mathbf{V} \mid \times \mid \mathbf{Z} \mid$ matrix whose z-th column is $[\pi_{(v,w)z}, v \in \mathbf{V}]$. Let $\mathbf{C}_w^* \subset R^{|\mathbf{z}|}$ denote the set of solutions $\xi \in R^{|\mathbf{z}|}$ to the system of linear equations

$$\mathrm{E}(Y \mid V = v, W = w) = \sum_{z \in \mathbf{Z}} \pi_{(v,w)z} \, \xi_z, \qquad \forall \; v \in \mathbf{V}. \tag{13}$$

Cross and Manski show that $\mathbf{H}_w^* \subset \mathbf{C}_w^*$. Thus, $\mathbf{C}_w^*$ is an outer identification region.

Suppose that the matrix $\mathbf{\Pi}$ has rank $\mid \mathbf{Z} \mid$. Then the system of equations (13) has either one solution or none at all. If it has one solution, $\mathbf{C}_w^*$ is a singleton, and $\mathbf{H}_w^* = \mathbf{C}_w^*$. Thus, assumption (9) yields point identification when $\mathbf{\Pi}$ has rank $\mid \mathbf{Z} \mid$. In the 2×2 case, $\mathbf{C}_w^*$ is the solution to the ecological inference problem developed by Goodman. If (13) has no solutions, then this is a useful diagnostic, and allows one to conclude that assumption (9) is incorrect.

2.3 Structural Prediction

Our discussion thus far may seem highly conceptual and inclined toward intellectual curiosity rather than practical application. However, the abstract study of the ecological inference problem is useful for its insights into structural prediction. For instance, political scientists often want to predict how an observed mean outcome, $\mathrm{E}(Y)$, say voting propensity, would change if the covariate distribution, were to change from $\mathrm{P}(X, Z)$ to some other distribution, say $\mathrm{P}^*(X, Z)$, that is from one set of racial contexts to another. It is common to address this prediction problem under the assumption that the mean regression $\mathrm{E}(Y \mid X, Z)$ is *structural*, in the sense that this regression would remain invariant under the hypothesized change in the covariate distribution. Given this assumption, the mean outcome under covariate distribution $\mathrm{P}^*(X, Z)$ would be

$$\mathrm{E}^*(Y) \equiv \sum_{x \in \mathbf{X}} \sum_{z \in \mathbf{Z}} \mathrm{E}(Y \mid X = x, Z = z) \, \mathrm{P}^*(Z = z \mid X = x) \, \mathrm{P}^*(X = x).$$

Suppose one wants to compute $\mathrm{E}^*(Y)$ and compare it to $\mathrm{E}(Y)$. A glaring barrier is that $\mathrm{E}(Y \mid X, Z)$ is not point-identified because of the ecological inference problem, and so our understanding of $\mathrm{E}^*(Y)$ is limited. A theme that has resonated throughout this chapter is that "limits" are not always as constraining as they may first appear, and that knowing what the limits are can be illuminating in and of itself because they reveal what is based on the known data and clear logic and what lies outside those realms. Our discussion and the findings summarized in Sections 2.1 and 2.2 show what one can learn about $\mathrm{E}^*(Y)$. For example, using the data alone, one can conclude that $\mathrm{E}^*(Y)$ lies in the set

$$\left\{ \sum_{x \in \mathbf{X}} \sum_{z \in \mathbf{Z}} \xi_{xz} \; \mathrm{P}^*(Z = z \mid X = x) \, \mathrm{P}^*(X = x); \; (\xi_{xz}, z \in \mathbf{Z}) \in \mathbf{H} \left[\mathrm{E}(Y \mid X = x, Z) \right], x \in \mathbf{X} \right\}. \tag{14}$$

This type of knowledge is helpful particularly for its ability to separate findings based on assumptions and findings that are necessarily true based on data constraints.

To illustrate structural prediction using the data alone, we summarize elements of an empirical application reported in Cross and Manski (1999), a working paper version of their 2002 article. They posed and addressed the counterfactual question: "What would be the outcome of the 1996 U.S. Presidential election if the U.S. population had a different composition, ceteris paribus?" To formalize the question, let x denote a particular state in the U.S. or the District of Columbia. Let Z denote attributes of individual voters possibly associated with voting behavior (e.g., age or race) in state x. Let $\mathbf{Y} = \{-1, 0, 1\}$ be the set of vote choices where $Y = 1$ if a person votes Democratic, $Y = -1$ if a Republican vote is cast, and $Y = 0$ otherwise (i.e. minor parties, abstentions, etc.).

In this setting $\mathrm{P}(Y \mid X = x)$ is the distribution of voting outcomes in state x, $\mathrm{P}(Z \mid X = x)$ is the distribution of voter attributes in the state, and $\mathrm{E}(Y \mid Z = z, X = x)$ is the Democratic plurality among voters in state x who have attributes z. Let S_x denote the number of electoral votes allocated to state x. Assuming there are no ties, the Democratic share of the electoral vote is $T \equiv \sum_{x \in \mathbf{X}} S_x \cdot \mathbf{1}[\mathrm{E}(Y \mid X = x) > 0]$ where $\mathbf{1}[\cdot]$ is the indicator function. The number of electors required to win the election is 270.

We know the outcome of the 1996 election, but would that outcome have differed if the composition of the population were different? That is, what would the outcome have been if the distribution of attributes in state x were $\mathrm{P}^*(Z \mid X = x)$ and its share of the Electoral College votes were S_x^*? To address this question, we maintain the key assumption that $\mathrm{E}(Y \mid \cdot, \cdot)$ is *invariant* in the sense that these conditional expectations remain unchanged under the hypothesized demographic change. This is a non-trivial assumption that hinges on the specification chosen for the covariates Z, but one that seems reasonable to entertain. To interpret this assumption, it may help to consider a behavioral model of the form $Y = f(X, Z, U)$ wherein vote choice is a function, f, of one's state of residence, X, personal attributes, Z, and other factors, U.

Then $\mathrm{E}(Y \mid \cdot, \cdot)$ is invariant if U is statistically independent of (X, Z) and if the distribution of U remains unchanged under the hypothesized demographic change.

If $\mathrm{E}(Y \mid \cdot, \cdot)$ is invariant, the predicted Democratic plurality in state x is

$$\mathrm{E}^*(Y \mid X = x) \equiv \sum_{z \in \mathbf{Z}} \mathrm{P}^*(Z = z \mid X = x)\, \mathrm{E}(Y \mid X = x, Z = z). \tag{15}$$

The predicted number of Democratic electoral votes is $T^* \equiv \sum_{x \in \mathbf{X}} S_x^* \times \mathbf{1}[\mathrm{E}^*(Y \mid X = x) > 0]$. These formulations make it clear that $[\mathrm{E}^*(Y \mid x), x \in \mathbf{X}]$ and thus T^* are functions of $\mathrm{E}(Y \mid \cdot, \cdot)$, which is unknown and an instance of the ecological inference problem. The identification region for $\mathrm{E}(Y \mid \cdot, \cdot)$ determines the region for T^* and provides the basis for evaluating our counterfactual.

We can obtain data and forecasts for demographic attributes to evaluate our counterfactual from the U.S. Census. In particular, we let our Z covariates be age (two categories: 18–54 years and 55+ years) and ethnicity (white, black and Hispanic).[3] Table 6 reports the bounds on $\mathrm{E}^*(Y \mid X)$ in 2004 and 2020. The table shows that the bounds on $\mathrm{E}^*(Y \mid X)$ in 2020 are wider than those in 2004 for all states. In 2004 there are 25 states where the bound on Democratic plurality is entirely a positive interval, and 11 states where the bound is entirely a negative interval. In 2020 the corresponding number of states is five and zero, respectively. The reason the bounds are wider in 2020 is simple. The forecast change in the distribution of demographic characteristics, $\mathrm{P}(Z \mid x)$, for each $x \in \mathbf{X}$ is more pronounced between 1996 and 2020 than between 1996 and 2004. The more $\mathrm{P}(Z \mid X = x)$ varies, the less information $\mathrm{P}(Y \mid X = x)$ conveys about $\mathrm{E}^*(Y \mid X = x)$.

From the bounds on $\mathrm{E}^*(Y \mid X = x)$ in a particular state, x, in 2004, we can predict the number of Electoral College seats the Democratic candidate will win in that state. For the 25 states where the bound on $\mathrm{E}^*(Y \mid X = x)$ is entirely a positive interval, we get the point prediction S_x^* as the number of seats won. And, for the 11 states where the bound on $\mathrm{E}^*(Y \mid X = x)$ is entirely a negative interval, our point prediction of the number of seats won is zero. In the remaining 15 states the bound on $\mathrm{E}^*(Y \mid X = x)$ straddles zero, and so we obtain no prediction for the number of Electoral College seats won by the Democratic candidate. In the absence of any cross-x or cross-state restrictions, we simply add these bounds, some of which reduce to a point, across all states to obtain the bound on T^*.

3 Conclusion

Long ago, Duncan and Davis (1953) pointed out the fundamental indeterminacy in ecological inference. Their simple analysis made clear that aggregate data only partially reveals the structure of individual behavior. Nevertheless, their contribution has largely been viewed as limited and an appreciation for the idea

[3] In this categorization, Hispanics are all Hispanics regardless of race, blacks are non-Hispanic blacks, and whites are all remaining non-Hispanics. Note also that while these are population figures, we treat them as citizen voting age population figures.

Table 6: **Bounds on $\mathrm{E}^*(y \mid x)$ and T^* in 2004 and 2020**

	1996	2004	2020
	$\mathrm{E}(y \mid x)$	Bound on $\mathrm{E}^*(y \mid x)$	Bound on $\mathrm{E}^*(y \mid x)$
Northeast			
Connecticut	0.102	[0.055 , 0.146]	[-0.053 , 0.252]
Maine	0.134	[0.103 , 0.168]	[-0.028 , 0.309]
Massachusetts	0.184	[0.141 , 0.223]	[0.025 , 0.346]
New Hampshire	0.057	[0.023 , 0.093]	[-0.116 , 0.237]
Rhode Island	0.171	[0.131 , 0.206]	[0.019 , 0.319]
Vermont	0.130	[0.091 , 0.172]	[-0.035 , 0.308]
New Jersey	0.091	[0.048 , 0.130]	[-0.056 , 0.231]
New York	0.134	[0.098 , 0.167]	[0.014 , 0.249]
Pennsylvania	0.045	[0.024 , 0.066]	[-0.068 , 0.162]
Midwest			
Illinois	0.086	[0.051 , 0.120]	[-0.050 , 0.222]
Indiana	-0.027	[-0.055 , -0.001]	[-0.158 , 0.098]
Michigan	0.072	[0.042 , 0.104]	[-0.064 , 0.218]
Ohio	0.035	[0.007 , 0.063]	[-0.097 , 0.171]
Wisconsin	0.060	[0.028 , 0.091]	[-0.093 , 0.221]
Iowa	0.060	[0.034 , 0.086]	[-0.078 , 0.206]
Kansas	-0.103	[-0.134 , -0.072]	[-0.262 , 0.046]
Minnesota	0.104	[0.068 , 0.140]	[-0.073 , 0.290]
Missouri	0.034	[0.013 , 0.056]	[-0.099 , 0.172]
Nebraska	-0.105	[-0.134 , -0.077]	[-0.256 , 0.032]
North Dakota	-0.038	[-0.065 , -0.013]	[-0.183 , 0.100]
South Dakota	-0.021	[-0.034 , -0.008]	[-0.171 , 0.125]
South			
Delaware	0.076	[0.041 , 0.110]	[-0.079 , 0.237]
District of Columbia	0.331	[0.299 , 0.364]	[0.253 , 0.401]
Florida	0.028	[-0.025 , 0.078]	[-0.157 , 0.209]
Georgia	-0.005	[-0.044 , 0.033]	[-0.166 , 0.155]
Maryland	0.075	[0.030 , 0.117]	[-0.085 , 0.233]
North Carolina	-0.022	[-0.060 , 0.016]	[-0.183 , 0.135]
South Carolina	-0.024	[-0.066 , 0.016]	[-0.175 , 0.120]
Virginia	-0.009	[-0.058 , 0.039]	[-0.172 , 0.152]
West Virginia	0.066	[0.035 , 0.101]	[-0.057 , 0.203]
Alabama	-0.033	[-0.067 , -0.001]	[-0.176 , 0.102]
Kentucky	0.005	[-0.028 , 0.038]	[-0.143 , 0.153]
Mississippi	-0.023	[-0.055 , 0.008]	[-0.164 , 0.113]
Tennessee	0.011	[-0.020 , 0.043]	[-0.135 , 0.160]
Arkansas	0.080	[0.048 , 0.117]	[-0.061 , 0.239]
Louisiana	0.069	[0.022 , 0.117]	[-0.104 , 0.247]
Oklahoma	-0.039	[-0.079 , 0.000]	[-0.191 , 0.109]
Texas	-0.020	[-0.058 , 0.018]	[-0.168 , 0.128]
West			
Arizona	0.010	[-0.039 , 0.059]	[-0.166 , 0.186]
Colorado	-0.007	[-0.068 , 0.053]	[-0.224 , 0.208]
Idaho	-0.108	[-0.159 , -0.060]	[-0.320 , 0.080]
Montana	-0.018	[-0.070 , 0.033]	[-0.231 , 0.190]
Nevada	0.004	[-0.053 , 0.061]	[-0.186 , 0.194]
New Mexico	0.034	[-0.001 , 0.068]	[-0.116 , 0.184]
Utah	-0.106	[-0.145 , -0.074]	[-0.284 , 0.046]
Wyoming	-0.078	[-0.133 , -0.028]	[-0.284 , 0.109]
Alaska	-0.100	[-0.155 , -0.039]	[-0.232 , 0.050]
California	0.057	[0.003 , 0.114]	[-0.085 , 0.200]
Hawaii	0.103	[0.079 , 0.130]	[0.028 , 0.189]
Oregon	0.047	[-0.014 , 0.110]	[-0.156 , 0.260]
Washington	0.069	[0.017 , 0.125]	[-0.116 , 0.272]
Democratic Electoral Votes, T^*	379	[302 , 477]	[51 , 538]

of bounds or an identification region has yet to fully emerge. Instead, the bulk of the effort has been directed toward obtaining point estimates or to somehow point identify a problem that is only partially identified. However, empirical researchers should be aware that no solution comes free. Every method yielding point estimates necessarily rests on assumptions that are strong enough to remove the indeterminacy of ecological inference. Researchers contemplating application of any method should carefully consider whether the associated assumptions are credible in their applications.

In our experience, it is rare in practice to find point-identifying assumptions that one can accept with great confidence. The prudent researchers, then, should resist the temptation to embrace any particular estimation method. Instead, the analysis of aggregate data should be a process. First, one should determine what one can learn from the data alone and without imposing any assumptions. Then, one should consider various assumptions that have identifying power. A productive approach is to "layer" the assumptions, imposing them sequentially in order of decreasing plausibility. As more assumptions are imposed, one will be able to draw conclusions that are increasingly sharp but decreasingly believable. This process of inference illuminates the respective roles that data and assumptions play in empirical research. Moreover, it enables both researchers and their consumers to adjudicate how best to reconcile the inevitable tension between the strength of conclusions and their credibility.

References

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

Allport, Floyd H. 1924. "The Group Fallacy in Relation to Social Science." *American Journal of Sociology* 29: 688–703.

Anselin, Luc and Wendy K. Tam Cho. 2002. "Spatial Effects and Ecological Inference." *Political Analysis* 10, 3 (Summer): 276–297.

Brown, P. J. and C. D. Payne. 1986. "Aggregate Data, Ecological Regression, and Voting Transitions." *Journal of the American Statistical Association* 81: 452–460.

Calvo, Ernesto and Marcelo Escolar. 2003. "The Local Voter: A Geographically Weighted Approach to Ecological Inference." *American Journal of Political Science* 47, 1 (January): 189–204.

Cho, Wendy K. Tam. 1998. "Iff the Assumption Fits...: A Comment on the King Ecological Inference Solution." *Political Analysis* 7: 143–163.

Cho, Wendy K. Tam and Brian J. Gaines. 2004. "The Limits of Ecological Inference: The Case of Split-Ticket Voting." *American Journal of Political Science* 48, 1 (January): 152–171.

Cho, Wendy K. Tam, George G. Judge, and Brian J. Gaines. 2005. "Occam's Razor and Information Processing and Recovery with Aggregate Data." Working Paper.

Cho, Wendy K. Tam and Albert H. Yoon. 2001. "Strange Bedfellows: Politics, Courts, and Statistics: Statistical Expert Testimony in Voting Rights Cases." *Cornell Journal of Law and Public Policy* 10, 2 (Spring): 237–264.

Cleave, N., P.J. Brown, and C.D. Payne. 1995. "Evaluation of Methods for Ecological Inference." *Journal of the Royal Statistical Society, Series A* 158, 1: 55–72.

Crewe, Ivor and Clive Payne. 1976. "Another Game with Nature: An Ecological Regression Models of the British Two-Party Vote Ratio in 1970." *British Journal of Political Science* 6, 1 (January): 43–81.

Cross, Philip J. and Charles F. Manski. 1999. "Regressions, Short and Long," presented at the 2000 Econometric Society World Congress, `http://ideas.repec.org/p/ecm/wc2000/0385.html`

Cross, Philip J. and Charles F. Manski. 2002. "Regressions, Short and Long." *Econometrica* 70 (1): 357–368.

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

Firebaugh, Glenn. 1978. "A Rule for Inferring Individual-Level Relationships from Aggregate Data." *American Sociological Review* 43, 4 (August): 557–572.

Freedman, D., S. Klein, J. Sacks, C. Smyth, and C. Everett. 1991. "Ecological Regression and Voting Rights." *Evaluation Review* 15, 6 (December): 673–711.

Freedman, D.A., S.P. Klein, M. Ostland, and M.R. Roberts. 1998. "Review of A Solution to the Ecological Inference Problem." *Journal of the American Statistical Association* 93, 444 (December): 1518–1522.

Freedman, D.A., M. Ostland, M.R. Roberts, and S.P. Klein. 1999. "Response to King's Comments" *Journal of the American Statistical Association* 94, 445 (March): 355–357.

Gehlke, C. E., and Katherine Biehl. 1934. "Certain Effects of Grouping upon Size of the Correlation Coefficient in Census Tract Material." *Journal of the American Statistical Association Supplement* 29: 663–664.

Good, I. J. 1963. "Maximum Entropy for Hypothesis Formation, Especially for Multidimensional Contingency Tables." *Annals of Mathematical Statistics* 64: 911–934.

Goodman, Leo A., 1953. "Ecological Regressions and Behavior of Individuals." *American Sociological Review* 18, 6 (December): 663–664.

Goodman, Leo A., 1959. "Some Alternatives to Ecological Correlation." *American Journal of Sociology* 64, 6 (May): 610–625.

Grofman, Bernard, Michael Migalski, and Nicholas Noviello. 1985. "The 'Totality of Circumstances' Test in Section 2 of the 1982 Extension of the Voting Rights Act: A Social Science Perspective." *Law and Policy* 7: 209–223.

Hammond, John L. 1973. "Two Sources of Error in Ecological Correlations." *American Sociological Review* 38, 6 (December): 764–777.

Hanushek, E., J. Jackson, and J. Kain. 1974. "Model Specification, Use of Aggregate Data, and the Ecological Correlation Fallacy." *Political Methodology* 89–107.

Hawkes, A. G. 1969. "An Approach to the Analysis of Electoral Swing." *Journal of the Royal Statistical Association, Series A* 132: 68–79.

Imai, Kosuke and Ying Lu. 2005. "Parametric and Nonparametric Bayesian Models for Ecological Inference in 2 × 2 Tables." Working paper. Princeton University.

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

Kleppner, Paul. 1985. *Chicago Divided: The Making of a Black Mayor*. DeKalb, IL: Northern Illinois University Press.

Kousser, J. Morgan. 1973. "Ecological Regression and the Analysis of past Politics." *Journal of Interdisciplinary History* 4, 2 (Autumn): 237–262.

Jones, E. Terrence. 1972. "Ecological Inference and Electoral Analysis." *Journal of Interdisciplinary History* 2, 3 (Winter): 249–262.

Lichtman, Allan J. 1974. "Correlation, Regression, and the Ecological Fallacy: A Critique." *Journal of Interdisciplinary History* 4, 3 (Winter): 417–433.

Loewen, James. 1982. *Social Science in the Courtroom: Statistical Techniques and Research Methods for Winning Class-Action Suits*. Lexington, MA: Lexington Books.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press.

Manski, Charles F. 2003. *Partial Identification of Probability Distributions*. New York: Springer-Verlag.

McCue, Kenneth F. 2001. "The Statistical Foundations of the EI Method." *The American Statistician* 55, 2 (May): 102–105.

Ogburn, William F., and Inez Goltra. 1919. "How Women Vote: A Study of an Election in Portland, Oregon." *Political Science Quarterly* 34: 413–433.

Openshaw, Stan and Peter Taylor. 1979. "A Million or so Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem." In *Statistical Applications in the Spatial Sciences,* ed. N. Wrigley. London: Pion. Pp 127–144.

Prais, S. J. and J. Aitchison. 1954. "The Grouping of Observations in Regression Analysis." *Review of the International Institute of Statistics* 22: 1–22.

Rivers, Douglas. 1998. "Book Review: A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data." *American Political Science Review* 92, 2 (June): 442–443.

Robinson, W.S. 1950. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review* 15, 3 (January): 351–357.

Salway, Ruth and Jonathan Wakefield. 2004. "A Comparison of Approaches to Ecological Inference in Epidemiology, Political Science, and Sociology." In *Ecological Inference: New Methodological Strategies*, King, Gary, Ori Rosen, and Martin Tanner, eds. Cambridge University Press. Pp. 303–332.

Simpson, E. H. 1951. "The Interpretation of Interaction in Contingency Tables." *Journal of the Royal Statistical Society, Series B* 13: 238–241.

Thomsen, S. R. 1987. *Danish Elections 1920–1979: A Logit Approach to Ecological Analysis and Inference.* Arhus: Politica.

Wakefield, Jonathan. 2003. "Sensitivity Analyses for Ecological Regress." *Biometrics* 59: 9–17.

Yule, G. Udny and M. G. Kendall. 1950. *An Introduction to the Theory of Statistics*. London: Griffin.