The Limits of Ecological Inference: The Case of Split-Ticket Voting

Wendy K. Tam Cho University of Illinois at Urbana-Champaign
Brian J. Gaines University of Illinois at Urbana-Champaign

We examine the limits of ecological inference methods by focusing on the case of split-ticket voting. Burden and Kimball (1998) report that, by using the King estimation procedure for inferring individual-level behavior from aggregate data, they are the first to produce accurate estimates of split-ticket voting rates in congressional districts. However, a closer examination of their data reveals that a satisfactory analysis of this problem is more complex than may initially appear. We show that the estimation technique is highly suspect in general and especially unhelpful with their particular data.

A large class of interesting problems in political science involves drawing inferences about the behavior of individuals using only aggregate data. Using data in which some information has been lost in the aggregation process to retrieve information about individuals is known as "ecological inference" or "cross-level inference." Such inferences are particularly important for the study of voting, since the use of secret ballots usually makes it impossible to obtain objective individual-level data on vote choices. Empirical work on voting behavior has thus had to proceed on two tracks. One may analyze respondents' self-reports of voting behavior (and other traits) from surveys, or else one may use aggregate official election returns (and other aggregated data) for some geographic area such as a district, constituency, precinct, county, commune, etc. Using survey data is, of course, not always an option and is rarely possible for most historical analyses. An additional complication is that even when survey data exist, they are prone to various biases and errors related to sample-selection effects, priming, systematic misreporting, and so on. In some cases, aggregate data may even provide better leverage on particular theoretical problems than do microlevel survey data, even when the behavior of interest is, in fact, actions taken by individuals (e.g., Kramer 1983). Analyzing aggregate data is often the only option.

Developing statistical methods for drawing useful estimates of individual-level behavior (e.g., voting) from aggregate data is, unfortunately, justly characterized as difficult. In ecological regression problems, we are interested in the joint distribution of two or more variables, but we observe only the marginal distribution of each variable. There is, of course, no unique "solution," since many different joint distributions are consistent with the observed marginals. This inverse problem is clearly ill-posed. Under these circumstances, one approach is to make a number of convenient assumptions to induce a completely specified statistical model that is amenable to standard estimation techniques. Though there are an infinite number of possible solutions, it is frequently true in a particular situation that some criteria may be more reasonable than others.

For ecological inference to yield genuine insight, a number of circumstances must be met. When these conditions are met or when this type of analysis is the only recourse, ecological inference may be a reasonable research strategy, though extreme caution must always be exercised, at both the analysis and the interpretation stages.

Wendy K. Tam Cho is Associate Professor of Political Science and Statistics at the University of Illinois at Urbana-Champaign, 361 Lincoln Hall, 702 S. Wright Street, Urbana, IL 61801. Brian J. Gaines is Associate Professor of Political Science and an affiliate of the Institute of Government and Public Affairs at the University of Illinois at Urbana-Champaign, 361 Lincoln Hall, 702 S. Wright Street, Urbana, IL 61801.

We would like to thank Bear Braumoeller, Bruce Cain, Michael Caldwell, Lawrence Cho, Christophe Crombez, Susan Jellissen, Masaru Kohno, Jim Kuklinski, Walter Mebane, Peter Nardulli, Brian Sala, Jasjeet Sekhon, Paul Sniderman, three anonymous *AJPS* reviewers, and reviewers 1–10 at the *APSR* for helpful comments. Cho also thanks the National Science Foundation (Grant No. SBR–9806448) for research support.

American Journal of Political Science, Vol. 48, No. 1, January 2004, Pp. 152–171

ISSN 0092-5853

At minimum, three conditions are necessary (but far from sufficient). First, the data should appear to be amenable to ecological inference, i.e., there should be evidence that the aggregate data are "informative" about the microlevel process. We elaborate further on the term "informative" below. Second, there should be some evidence that the aggregation process did not introduce bias that is not modeled. And, third, one should have a good microtheory and an explicit understanding of how that microtheory should be related to the observed macro data. When these conditions hold, one *may* wish to implement an ecological inference model.

In this article, we discuss the circumstances surrounding this highly tenuous estimation. To situate our discussion, we focus on the EI estimator proposed by King (1997) and its application to the case of split-ticket voting, as expounded on by Burden and Kimball (1998). We proceed as follows. The next section identifies and describes three conditions necessary for ecological inference to be a useful method. Thereafter, we revisit each condition in more depth by reconsidering Burden and Kimball's anlysis of split-ticket voting in American elections. Our analysis indicates that the Burden and Kimball study yields little true insight into the split-ticket voting phenomena, and more generally, that their foray into the aggregate data realm is illustrative of the problems that one should expect to encounter when conducting ecological inference analyses. We conclude that while the challenges inherent in ecological inferences are not insurmountable, the careful researcher will find the task to be fiercely daunting.

Conditions Amenable to Ecological Inference

The first condition we explore is the idea that the aggregate data need be "informative" concerning the underlying microlevel data. As Robinson (1950) has shown, there is no clear or direct relationship between data that are observed at different levels of aggregation. Indeed, this conundrum has puzzled scholars for decades (Gehlke and Biehl 1934). Nonetheless, some aggregate data are more informative about the microdata than others. In this section, we focus on what it means for aggregate data to be "informative," and what consequences arise when data are not very informative, but one proceeds with estimation just the same. We then discuss how aggregation from the individual level can introduce troublesome biases. Lastly, we describe the role of microtheories in the analysis of macrodata. Without loss of generality, our discussion is couched in a framework where the macrolevel data are election districts and the microlevel counterparts are the individual-level voting data.

TABLE 1 A Reduced Split-Ticket Problem for District *i*

	House Vote		
Presidential Vote	**Democrat**	**Republican**	**Vote**
Democrat	β_i^b	$1 - \beta_i^b$	X_i^b
Republican	β_i^w	$1 - \beta_i^w$	X_i^w
Fraction of Voters	T_i	$1 - T_i$	1

Informative Data

Deterministic Information: Bounded Parameters

One way to gauge the level of information contained in aggregate data is to consider all of the deterministic information contained therein. Consider a simple problem in split-ticket voting like the one shown in Table 1. Data for each district can be summarized by such a table. The available data include the values T, the Democratic proportion of the House vote, $(1 - T)$, the Republican proportion of the House vote, X^b, the Democratic proportion of the Presidential vote, and $X^w = 1 - X^b$, the Republican proportion of the Presidential vote.[1] What we do not have, but may be interested in estimating, are the proportions of split-ticket votes: β^w, the proportion of all those who voted for the Republican presidential candidate who also voted for the Democratic House candidate; and $1 - \beta^b$, the proportion of those who voted for the Democratic presidential candidate who also voted for the Republican House candidate.[2] Since vote shares necessarily fall between 0 and 1, the unknown parameters β^b and β^w fully characterize the table. The extent to which these parameters are further bounded within the unit square is the deterministic aspect of an aggregate data set.

Suppose that the district shown in Table 1 had 100 voters, and that the Democratic vote totals for president and House were 60 and 30, respectively. In that case, there

[1] Here, we retain Burden and Kimball's notation. The superscripts b and w are mnemonics for "black" and "white," inapt to split-ticket voting, but left over from the main running example in King's book, race and voting.

[2] One way to estimate these parameters is through the OLS model originally proposed by Goodman (1953, 1959), where $T = \beta^b X^b + \beta^w X^w$. Rearranging terms gives us the more familiar slope-intercept form, $T = \beta^w + (\beta^b - \beta^w) X^b$.

cannot have been more than 30 voters who supported both Democrats, and β^b cannot exceed $\frac{30}{60} = 0.5$. Likewise, β^w has an upper bound of $\frac{30}{40} = 0.75$, while both parameters have a lower bound of 0. In this way, combinations of marginal totals may exclude some values for each parameter, for each district. Note that in distributing the 30 Democratic House votes between the Democratic and Republican presidential voters, we simultaneously determine both β^b and β^w, since the parameters are dependent. If $\beta^b = 0.5$, then β^w is necessarily 0, and so on.[3]

By plotting all logically possible pairs of parameter values, one can succinctly summarize the deterministic information for each observation. Since $\beta^w = \frac{T - \beta^b X^b}{X^w}$, when one plots the possible values of β^w on the y-axis and the values of β^b on the x-axis, the result is a line with intercept $\frac{T}{X^w}$ and slope $-\frac{X^b}{X^w}$. This line has been termed a "tomography line," and there is one for each observation.[4] Both parameters are bounded within the [0, 1] interval, but those lines that do not extend across the entire unit square are further bounded, and one may be more successful when estimating the true parameter values for those observations. For estimation problems that can be simplified to 2 × 2 tables, then, a "tomography plot" succinctly displays the scope of the problem.

Qualitative Assessment of Nondeterministic Information

In addition to taking account of the deterministic bounds, one might incorporate some kind of assumption about how districts are related in order to arrive at estimates of plausible mean parameter values for a set of districts, or, sometimes, of parameters for each district. There are thus two diagnostic uses for tomography plots. First, they show all available deterministic information in a problem, and thereby reveal, in an informal sense, how constrained are the parameters, and thus how easy or hard the estimation problem will be. Second, one may examine these plots to assess whether an assumption that the (β^b, β^w) pairs were drawn from a distribution with a known form seems reasonable for the data at hand. The simplest distributional assumption is the case where all the β_i^ws are equal and all the β_i^bs are equal. In this case, it is easy to determine the values of the common β^w and β^b. Consider the very simple case with two observations or districts. The subscripts indicate the district.

$$T_1 = X_1\beta^b + (1 - X_1)\beta^w \tag{1}$$

$$T_2 = X_2\beta^b + (1 - X_2)\beta^w. \tag{2}$$

In this case, one can easily solve equation (1) for β^b to obtain $\beta^b = \frac{T_1 - (1 - X_1)\beta^w}{X_1}$. Since β^b has a common value across districts, we can then substitute this value for β^b into equation (2) to obtain a value for β^w, $\beta^w = \frac{T_2 X_1 - T_1 X_2}{X_1 - X_2}$. Likewise, by solving for β^w first and then substituting that expression back into the original equation, one finds that $\beta^b = \frac{T_2(1 - X_1) - T_1(1 - X_2)}{X_2 - X_1}$. Note that in this case all of the tomography lines will intersect at a common point—the common value of β^w and β^b. However, virtually all tomography plots are inconsistent with a single point of intersection, and instead, imply many different points of intersection. The obvious explanation in these cases is that all of the β^ws and β^bs are not equal to one another. That is, from precinct to precinct, the proportion of people splitting tickets varies. One way to proceed is to make some assumption about the underlying joint distribution for (β^b, β^w). King's model imposes the assumption that the joint distribution of β^b and β^w is truncated bivariate normal. Hence, when implementing King's model, one examines tomography plots for some evidence of consistency with an underlying truncated bivariate normal (TBVN) distribution.[5]

Testing a hypothesis about the distribution from which data were drawn is fairly straightforward when the data are directly observed. In this case, however, we have only a range of possible values for each observation, as mapped out by the lines. A single tomography plot is consistent with many different individual-level data sets, so many different joint distributions will be consistent with any given set of tomography lines. In that sense, the information gleaned from tomography plots is never more than suggestive and does not allow one to make definitive claims about whether particular distributional assumptions obtain. To say that a tomography plot is "informative" is merely to report that one or two conditions are met. If most of the tomography lines seem to intersect in a region, then it is more likely (but not certain) that the actual individual-level data are clustered there. In turn, this area marks a plausible location for the mode of the joint distribution of βs. Second, if there are relatively narrow bounds on one or both parameters, one can further

[3] Duncan and Davis (1953) is the canonical source on how to compute upper and lower limits (bounds) on the possible parameter values in light of the known marginals.

[4] Achen and Shively (1995, 208–09) originally proposed that one can succinctly summarize all the known information in an aggregate data problem by creating a plot with a line for each of the observations. King (1997) later applied the name "tomography plot."

[5] King also has a nonparametric model, but it is infrequently used. Indeed, we have never seen an application of it, by King or anyone else. Hence, we do not discuss this model hereafter.

TABLE 2 The Link Between Tomography Plots and the Distributional Assumption

	TBVN Assumption Correct	TBVN Assumption Incorrect
	Cell 1	*Cell 2*
"Informative"	*Correct* standard errors	*Incorrect* standard errors
Tomography Plot	Small standard errors	Small and misleading standard errors
	Cell 3	*Cell 4*
"Uninformative"	*Correct* standard errors	*Incorrect* standard errors
Tomography Plot	Large standard errors	Large and misleading standard errors

limit the possible parameters of this distribution. At best, though, one can conclude that the data are consistent with a unimodal distribution, when there is an area of intersection. On the other hand, if no area of intersection is evident and the bounds are wide, the implication is that the TBVN distributional assumption is not reasonable.[6] Whether the distributional assumption seems to hold or not, meanwhile, is important not only for the purposes of estimating means, but also because, at the estimation stage, the computation of the standard errors is based on the distributional assumption. So whether the standard errors are correct or incorrect also depends on whether the distributional assumption is correct or incorrect. This logic is summarized in Table 2.

King contends that an "informative" tomography plot can reasonably be assumed to have been generated by a truncated bivariate normal distribution. That is, he would attribute a higher probability that the output from data analysis is summarized by Cell 1 rather than by Cell 2. Similarly, if a tomography plot is uninformative, he claims the data are less likely to have been generated from a TBVN, and the situation is more likely to be summarized by Cell 4 than by Cell 3. There is no particular reason to believe that the diagonal cells in Table 2 are more likely than the off-diagonal cells. King's contention here amounts to an a priori assumption. Indeed, it would be very hard to make a formal probabilistic argument about this link. Our examples that follow should produce more intuition on what is and is not revealed by a tomography plot. At best, a researcher hopes that the tomography plot will be informative: if it is not, the resulting standard errors may be too large to be useful, or simply incorrect (see King 1997, ch. 16).

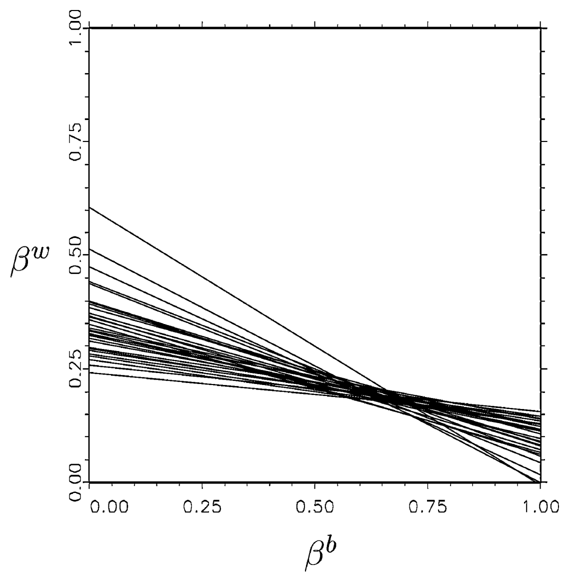
One's assessment of whether the distributional assumption is correct thus depends on the nature of the tomography plot, though, of course, this assessment is never definitive. Moreover, deciding whether a tomography plot is informative is something of an art, no one has devised a concrete measure for "informativeness" or any formal test for accepting or rejecting the TBVN distributional assumption (or any other distributional assumption) on the basis of the plot.

Consider Figure 1. By the reasoning just discussed, this plot is informative. First, while the bounds on β^b span the entire permissible [0, 1] range, the bounds on β^w are more narrow, and thus limit the range of possible true values. Second, there is a general area of intersection of tomography lines. If these lines are related (as implied by the distributional assumption in the EI model) then the true points on each line should fall within the area where the lines generally intersect. In this plot, the area of "general intersection" clearly falls at approximately $(\beta^b, \beta^w) = (0.65, 0.20)$. While this point may not represent the true values for β^b and β^w for all districts, if we have no other information, these values seem to be reasonable first guesses. Of course, not all tomography plots are as seemingly informative. Sometimes the bounds will not be very informative at all, and, in addition, the tomography lines will not display any sort of commonality. Obviously, in such cases, it is far riskier to force the distributional assumption on the data.

Contrast the tomography plot in Figure 1 with the tomography plot on the left in Figure 2, where we have very little deterministic information about the underlying data. No "general area of intersection" seems to be present, and the bounds on both parameters are very wide. If one insists on proceeding with EI, it will impose a truncated bivariate normal distribution and then produce estimates for the β parameters (overall means and one per district) accordingly. The mode of the assumed TBVN, indicated by the small square in the plot on the right in Figure 2, is the estimate for the mean β^b and β^w values for these data. However, "if the ultimate conditional distributions are not reasonably close approximations to the

[6] If the tomography plot leads one to reject the TBVN distributional assumption, a model incorporating a TBVN distribution might still be adequate provided that one conditions on appropriate covariates. If the data are consistent with different TBVNs, conditional on values of some set of covariates, then the difficulty for estimation is model specification. In this sense, the tomography plot can be thought of as a diagnostic for the necessity of adding covariates to the model.

FIGURE 1 Informative Tomography Plot



This tomography plot is informative for two reasons. First, all of the lines intersect in one general area of the plot. This gives us some confidence in the assumption that all of the lines are related—a key assumption of the EI model. Second, while the bounds on β^b are wide, the bounds on β^w are relatively narrow.

truth, incorrect inferences may result" (King 1997, 185). Here, the tomography plot has not given us a good indication that the distributional assumption is correct—quite the contrary.[7]

Moreover, it is perhaps even more important to acknowledge that cross-level inference is always tenuous. In particular, a plot may appear to be informative even if the underlying data generation process is not at all well approximated by a TBVN. On the other hand, a plot may not appear to be informative even though the true parameters describing behavior do conform well to a TBVN distribution. A variety of possible situations are illustrated in Figure 3. In each plot, the true (β^b, β^w) pairs for each district are indicated by a point on the tomography line. In the first plot, the parameters are drawn from a TBVN distribution, and the plot is properly informative. In this case, a researcher would likely proceed properly based on this diagnostic. In the second plot, the parameters are not drawn from a TBVN distribution, but the tomography lines nonetheless appear to suggest a mode (i.e., it appears to be an "informative" plot).[8] Here, a researcher who proceeded with confidence would be grossly misled, and the analysis would suffer accordingly. In the third plot, the parameters are drawn from a TBVN distribution, but it has relatively large standard deviations on both the β^b and β^w parameters, and the resulting tomography plot does not appear at all "informative." On the basis of such a tomography plot, there is no reason to favor the truncated normal distribution as the underlying distribution, even though, in this instance, it happens to be correct. In short, inspecting tomography plots is worthwhile because they illustrate bounds, but researchers must understand that they are not definitive with respect to distributional assumptions.

Aggregation Bias

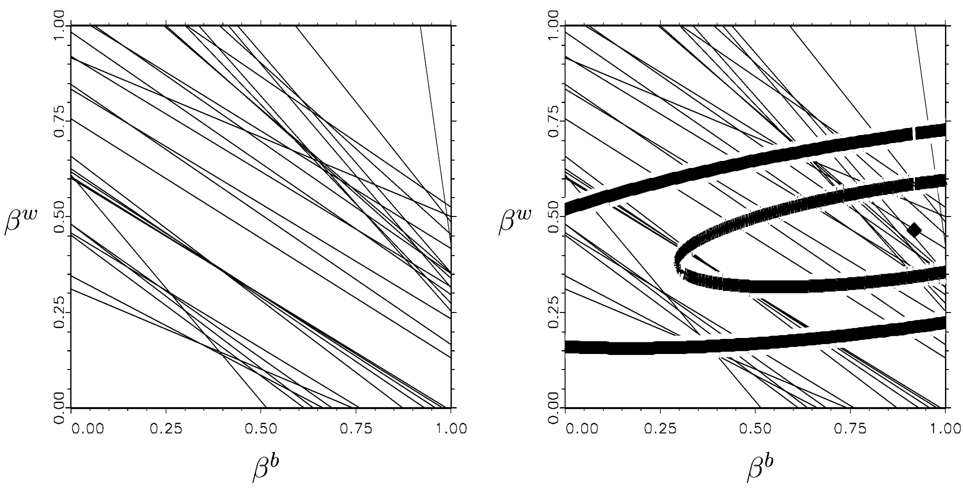
A second condition that helps to surmount the huge barriers to making ecological inferences is having data that aggregate without bias. It is usually possible to obtain reasonable estimates of individual-level parameters given only aggregated data if the aggregated data set contains no aggregation bias. The assumption of no aggregation bias holds if the parameters (β^b and β^w) are not correlated with the regressors, i.e., the *X* variable. In this application, that would mean that levels of Democratic President-Republican Representative and Republican President-Democratic Representative voting are not correlated with levels of support for the Presidential candidates. In fact, if no aggregation bias exists in the data, simple OLS will provide reasonable, unbiased, and consistent estimates of the overall means of the β parameters (Goodman 1953). EI should perform likewise. So in the very special case wherein data exhibit no aggregation bias, there is no reason to favor EI over OLS.[9] Meanwhile,

[7] It may be true that the distributional assumption will be reasonable for data despite the appearance of multiple modes in the tomography plot, because the *appearance* of multiple modes is, after all, subjectively assessed. This fact again emphasizes the limited utility of this diagnostic for determining whether a truncated bivariate normal distribution is a reasonable distribution for the data. In King's "Checklist" for ecological inference, Item 12b notes that even if EI fits a high variance TBVN (i.e., one with "very wide contours"), because of the presence of what appear to be multiple modes, "the model should probably be modified to fit this feature of the data [the multiple modes] anyway" (King 1997, 284).

[8] The parameters for this plot were chosen via a procedure described in King (1997, 162), not from an explicit distribution.

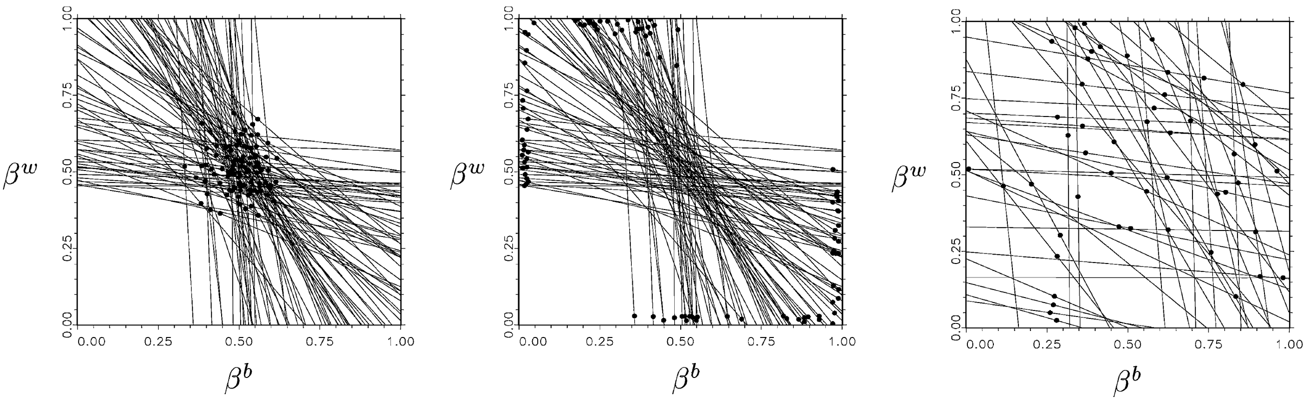
[9] One might choose EI over OLS because EI will output estimated β parameters for every district. Burden and Kimball, for instance, sought to test hypotheses about district-level variance in ticket-splitting, and so needed district-level estimates. This ostensible "advantage" of EI, however, is largely illusory, as these district estimates are not consistent and may lead to erroneous inferences. Some of the grave risks entailed in using EI district estimates as dependent variables in other models have been chronicled by Herron and Shotts (2003, 2004) and McCue (2001).

FIGURE 2 Uninformative Tomography Plot



These tomography plots are much less informative than the tomography plot in Figure 1. The lines do not intersect in any one general area of the plot. In addition, the bounds on both β^b and β^w are very wide and span virtually the entire permissible range. The elliptical segments on the right are contour lines that represent the estimated truncated bivariate normal distribution.

FIGURE 3 A Panel of Tomography Plots



Each plot indicates one of many situations that may characterize a tomography plot. Tomography plots can be helpful diagnostics, but are highly indeterminate all the same.

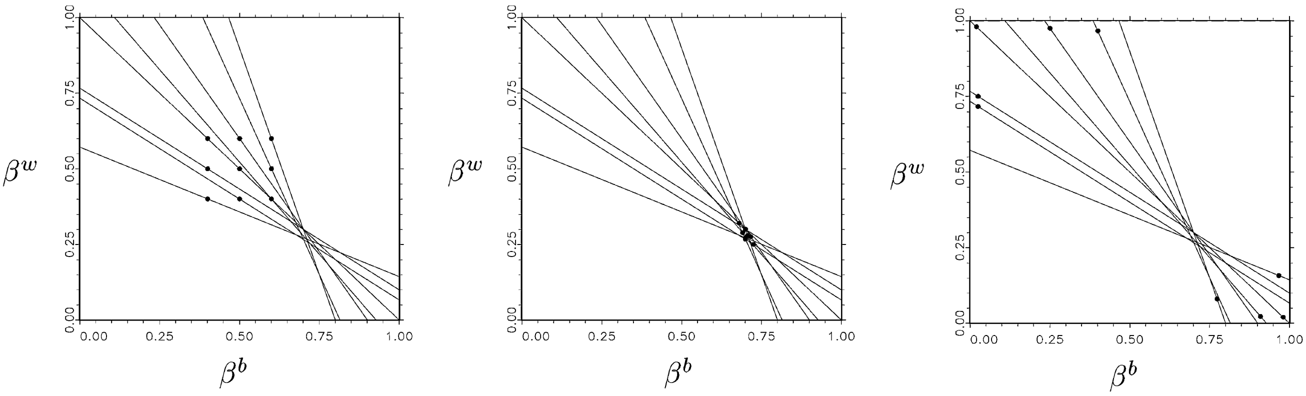
when data are affected by aggregation bias, neither model is trustworthy.[10]

Figure 4 uses tomography plots to illustrate how aggregation bias causes difficulties for ecological inference. All panels show nine districts, each having 100 voters, with β^b and β^w representing the proportions voting straight Democratic and voting for the Republican presidential candidate and the Democratic House candidate, respectively. In each scenario, these values are known, but we consider how the analyst not knowing them would proceed.

If the true voting patterns are represented by the left panel, the tomography plot is clearly misleading. The lines intersect around (0.7, 0.3), but this is not a good estimate for the β pairs, whose actual mean is (0.5, 0.5). The source of the error is aggregation bias: both β parameters are highly positively correlated with X, so that as one moves up and right on the plot, the slopes of the tomography lines (which are entirely determined by X) decrease, causing the misleading region of intersection in the lines.

[10] Indeed, in one Monte Carlo analysis, the correlation between EI and OLS estimates was 0.98 (Cho and Yoon 2001).

FIGURE 4 Aggregation Bias and Tomography Plots



The left plot illustrates how correlation between the true β parameters and the regressor causes bias in β estimates. Both the center and right panels have very little aggregation bias (correlations below 0.10). The TBVN distributional assumption fits only the center panel. In the right panel, the TBVN distribution imposed by EI yields incorrect, biased estimates.

Of course, the difficulty in a real data-analysis situation, in which one does not know the true β values, is that there are so many possible scatters of (β^b, β^w) pairs for a given set of tomography lines. In this artificially simple problem, wherein very few districts each have few voters, there are about 2×10^{14} different possible joint distributions of β^b and β^w arising from the known vote totals.[11] Knowing only the aggregates and their matching tomography lines, one cannot distinguish between the situations portrayed in the left, center, and right panels. It is thus very difficult to glean any information about aggregation bias directly from a tomography plot, except in the artificial situation wherein the true β values are known. Furthermore, the contrast between the middle and right panels (analogs to the left and middle panels of Figure 3) demonstrates that even very low aggregation bias does not guarantee that an informative tomography plot will not be misleading. In both cases, the correlations between both β parameters and X are less than 0.10 in absolute value, but only in the center case is the tomography plot correctly informative. In the right panel, the mean of β^b is 0.48, and the mean of β^w is 0.52 notwithstanding the intersection of lines in the vicinity of (0.7, 0.3). Assuming a TBVN centered there will, of course, result in faulty estimates.

To get a better sense for the degree to which aggregation bias foils ecological inference, consider next the results from a Monte Carlo simulation displayed in Figure 5. Here, data were constructed to exhibit aggregation bias but to be consistent with the distributional and spatial autocorrelation assumptions of the EI model.[12] In this simulation, 250 data sets were generated exactly according to the description in King (1997, 161). The data were drawn from a TBVN having parameters $\beta^b = \beta^w = 0.5$, $\sigma_b = 0.4$, $\sigma_w = 0.1$, and $\rho = 0.2$.[13] The true values, $\beta^b = \beta^w = 0.5$, are marked in the plots by a vertical line. For each simulation, we have drawn a bar centered on the point estimate for the parameter and model in question, extending one estimated standard error to each side. The error bars in Figure 5 clearly indicate that, even accounting for the standard errors, the estimates are inaccurate.[14] The sense of precision is overstated more by EI than OLS. On average, EI's estimates for β^b are 25 S.E.s from the true value, and its estimates of β^w are -14.7 S.E.s from the true value. In the OLS model, meanwhile, β^b is 18.8 S.E.s from the true value while β^w is -11.4 S.E.s from the true value, on average. Obviously, the standard errors are erroneously estimated and suggest more precision than

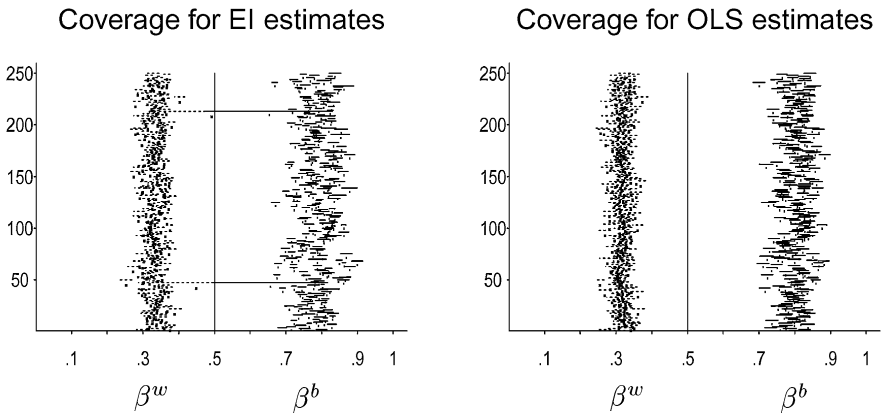
[11] We assume districts are distinguishable. The magnitude of this number clarifies why it is customary to treat β parameters as continuous variables and tomography lines as actual lines, not sets of points. Although actual ecological inference problems are discrete, there is no danger in studying continuous approximations except for very small or very highly bounded problems.

[12] The EI model incorporates three assumptions: the distributional assumption; the assumption of no spatial autocorrelation; and the assumption of no aggregation bias. For an extended discussion of the model assumptions, see King (1997) and Cho (1998).

[13] The mean correlation between X^b and the parameters in the 250 data sets is 0.69. The minimum correlation obtained via King's procedure was 0.21, while the maximum correlation was 0.97.

[14] There are two instances out of 250 simulations where EI produced extremely wide error bars that actually reach the true parameter values. However, these two cases are so anomalous that they seem likely to be the result of erroneous calculations by the *EzI* estimation program.

FIGURE 5 Consequence of Aggregation Bias



Pictured here are error bar plots from a Monte Carlo simulation with data which are consistent with the distributional and spatial autocorrelation assumptions but inconsistent with the aggregation bias assumption. The true parameter values are marked by the long vertical lines. The error bars to the left of the vertical line are for β^w. The error bars to the right of the vertical line are for β^b. Both β^b and β^w have a true parameter value of 0.5.

actually exists. Hence, even if the data are consistent with the other assumptions, if the parameters are correlated with the regressors, neither OLS nor EI will yield accurate results. Neither model is robust against aggregation bias.

The question of whether we are able to make reasonable ecological inferences turns on the issue of aggregation bias (Cho 1998). Though King claims that his method is "robust" to violations of the aggregation bias assumption, the evidence strongly suggests otherwise. King's claim originates in (and holds only for) an unorthodox definition of "robustness." He contends that EI is robust because it will never produce estimates of the β parameters which are outside the [0, 1] bounds. But estimates constrained to respect bounds need not be close to the truth, or even within a few standard errors of the actual values. Indeed, King's model does not produce unbiased or consistent estimates in the traditional statistical sense of those words when aggregation bias is present. When regressors are correlated with parameters, the estimates from EI are not equal to their respective population parameters, in expectation, and the discrepancy between the estimates and the true values does not converge in distribution to zero as the sample of data points becomes large (Cho 1998). Moreover, there is also evidence that the estimation of the standard errors is inaccurate as well. A fundamental issue for the split-ticket voting analyst, then, is whether aggregation bias exists in the election-returns data set. If rates of ticket splitting vary systematically according to how competitive the district was in the presidential race, then estimates from the King model (without covariates) will be untrustworthy.[15]

Microtheory

There is, no doubt, considerable variation in the precission of theory informing data analyses in the social sciences. Works aiming to test exact predictions from fully specified formal models are surely in the minority, and purely inductive exercises in which the authors cast about for relationships among a large number of variables that merely seem likely to be connected are not rare. We hesitate to take a strong position on the strict necessity of strong theory in all instances. However, in the case of ecological inference, we begin with the knowledge that aggregation can easily obscure data generating processes and microlevel mechanisms. Robinson's

[15] Violation of the spatial autocorrelation assumption has deleterious consequences as well. While violations of this assumption do not cause bias, they do affect the precision of the estimate. For an extensive discussion, see Anselin and Cho (2002). While Monte Carlo experiments permit exploration of the unique problems associated with each possible violation of EI model assumption, in real-world aggregate data it is typically the case that more than one assumption is violated (see King 1997, 159). Because there is not one problem but a whole host of potential problems, it is difficult to pinpoint the precise difficulty in any aggregate data analysis.

seminal work on ecological inference (1950) broached the highly memorable example of literacy and nonnativity. Upon discovering that states and regions with more foreign-born residents are, on average, more literate, one could infer that immigrants to America tended to be highly fluent in English. Even a casual acquaintance with American history, however, would suggest an alternative logic: immigrants tended to congregate in areas whose native-born populations were comparatively educated and literate. With the aggregate-level finding of a positive correlation in hand, one could work backwards to those rival accounts (and others), and not be in a position to choose one over the other on purely statistical grounds. With additional aggregated data, one could test their plausibility. But in the absence of additional data, it is prior knowledge and the credibility of the rival microlevel accounts that direct us to favor one account over the other. Our point, then, is less the purist's stance that theory must always precede empirical analysis than the common-sense argument that ecological inferences ought to be accompanied by an explicit microlevel theory.

To anticipate our arguments about voting, an analysis of ticket-splitting, or of transition probabilities from an earlier election to a later election, ought to be sensitive to the wealth of knowledge accrued about voting behavior and candidate strategy. Since it will always be the case that many alternative microlevel data generating processes could produce the same pattern of aggregate results, unambiguous ecological inferences are rare indeed, and strong claims of adjudication between rival models need to be very explicit about microlevel mechanisms. Furthermore, tests have to be constructed around the micrologic, with due attention to both (or all) rivals.

It is rarely ever simple to move from a microtheory to a macrolevel analysis, or from macro data analysis to a microlevel inference. Achen and Shively (1995) provide numerous examples where intuition goes wrong, and they demonstrate mathematically how relationships vanish or reverse in the process of aggregation. They describe proper macrolevel specification as "a subject with no simple relation to microlevel setups where our theories and intuitions apply" and assert that "macromodels will in general confound conventional statistical procedures" (1995, 95). Clearly, the upshot is that models that provide a good fit to the aggregate data may not provide an accurate portrayal of the underlying individual-level behavior. Indeed, this is the ecological fallacy—that what appears to be the case among macrounits may be vastly misleading with regard to the microunits. Explicit attention to microlevel theories, then, is important and should inform any analysis of aggregated data precisely because "reverse engineering" proves so vexing.

Case Study: Split-Ticket Voting in 1988

We now turn from this more theoretical discussion about the conditions under which ecological inference can be reasonable to an application of ecological inference techniques to split-ticket voting. In particular, we examine Burden and Kimball's (1998) analysis of ticket splitting at the Congressional district level. Their analysis of estimates based on King's (1997) EI method leads them to conclude that, contrary to previous findings (e.g., Alesina and Rosenthal 1995; Fiorina 1996), "voters are not intentionally splitting their tickets to produce divided government and moderate politics" (Burden and Kimball 1998, 533). Instead, they claim, ticket splitting is primarily the result of lopsided congressional campaigns in which well-funded, high-quality incumbents tend to run against unknown, underfunded challengers.

Our examination will show that there are two main reasons to doubt the generality and veracity of their conclusions. First, the EI model is not well-suited to these data. Neither of the previously discussed necessary conditions is met: these data are neither informative nor immune to aggregation bias. Second, their test is not informed by serious consideration of microlevel theories. There are, as well, a number of less fundamental difficulties with their analysis including an oversimplification of the full ticket splitting problem, an imperfect data set,[16] and use of a buggy EI software program.[17]

Our tactic is to revisit Burden and Kimball's analysis, highlighting the critical decisions at each stage, and focusing on how an ideal treatment of the split-ticket voting problem would differ. Our primary goal is to demonstrate that deriving insight into why individuals split their ballots by examining only aggregate data is far more difficult than their article implies. Because our main point is to focus on the extreme difficulty in deriving estimates of individual behavior using only aggregate data, we do not provide a revised, competing analysis of district-level split-ticket voting estimates. Instead, we precisely identify the major barriers to such an analysis and explicate what

[16] To maximize consistency, all of the estimates reported in this article have been obtained using the original Burden-Kimball data set.

[17] See the appendix for a brief discussion of this issue.

TABLE 3 The Complete Vermont Presidential-House Vote-Splitting Problem

	House of Representatives Vote								
Presidential Electors Vote	A	B	C	D	E	F	G	None	Total
0	v_{0A}	v_{0B}	v_{0C}	v_{0D}	v_{0E}	v_{0F}	v_{0G}	v_{0n}	124331
1	v_{1A}	v_{1B}	v_{1C}	v_{1D}	v_{1E}	v_{1F}	v_{1G}	v_{1n}	115775
2	v_{2A}	v_{2B}	v_{2C}	v_{2D}	v_{2E}	v_{2F}	v_{2G}	v_{2n}	1000
3	v_{3A}	v_{3B}	v_{3C}	v_{3D}	v_{3E}	v_{3F}	v_{3G}	v_{3n}	275
4	v_{4A}	v_{4B}	v_{4C}	v_{4D}	v_{4E}	v_{4F}	v_{4G}	v_{4n}	205
5	v_{5A}	v_{5B}	v_{5C}	v_{5D}	v_{5E}	v_{5F}	v_{5G}	v_{5n}	189
6	v_{6A}	v_{6B}	v_{6C}	v_{6D}	v_{6E}	v_{6F}	v_{6G}	v_{6n}	164
7	v_{7A}	v_{7B}	v_{7C}	v_{7D}	v_{7E}	v_{7F}	v_{7G}	v_{7n}	142
8	v_{8A}	v_{8B}	v_{8C}	v_{8D}	v_{8E}	v_{8F}	v_{8G}	v_{8n}	113
9	v_{9A}	v_{9B}	v_{9C}	v_{9D}	v_{9E}	v_{9F}	v_{9G}	v_{9n}	1134
None	v_{NA}	v_{NB}	v_{NC}	v_{ND}	v_{NE}	v_{NF}	v_{NG}	v_{Nn}	?
Total	98937	90026	45330	3110	1455	1070	203	?	≥243328

Presidential Candidates: 0. Republican (Bush); 1. Democrat (Dukakis); 2. Libertarian (Paul); 3. National Economic Recovery/Independent (Larouche) 4. New Alliance (Fulani); 5. Populist (Duke); 6. Peace and Freedom (Lewin); 7. Socialist/Liberty Union (Kenoyer); 8. Socialist Workers (Warren); 9. Scattering; A. Republican (Smith); B. Independent/Socialist (Sanders); C. Democrat (Poirier); D. Libertarian (Hedbor); E. Liberty Union (Diamondstone); F. Small is Beautiful (Earle); G. Scattering

remains to be done before accurate estimates can be produced. Accordingly, this article endeavors to delineate the conditions under which the EI model is an appropriate analytical tool.

Assessing the Split-Ticket Voting Data

Conceptualizing the Problem as a Multi-Stage Estimation

Burden and Kimball do not tackle American voting in all its complexity, but, rather, examine only two ticket-splitting scenarios. First, they analyze Presidential and House votes and then, separately, Presidential and Senate votes, ignoring House-Senate splits and all other races on the ballot. Thus, they consider not the very high dimensional problem of all varieties of ticket-splitting on complete ballots, but only two sets of two-way tables. Since EI is designed for 2 × 2 problems (i.e., it assumes dichotomous categorical variables), they further simplify the analysis by discarding all votes not cast for one of the two major parties and by assuming (falsely, as they recognize) that there are *no* ballots featuring choices in congressional contests but not choices in the presidential contest.[18] They thereby reduce the size of the table describing each House district to 2 × 3. Table 3 shows the full House-President case for Vermont, with each table entry, v_{ij}, representing a count of votes cast for presidential candidate i and House candidate j. Table 4 shows the simplified version analyzed in two steps by Burden and Kimball.[19]

In Table 1, we assumed a simple problem in which all voters had made choices for both Representative and President. Table 4, by contrast, acknowledges abstention from U.S. House voting. Burden and Kimball's Table A–1 is a general form of our Table 4 (except that it transforms the vote frequencies into row proportions). To submit data in this form to EI, Burden and Kimball summed across the first two columns to create another 2 × 2, with House-Vote and No-House-Vote for columns, and then estimated the Democratic and Republican presidential vote shares for only those voters who did not abstain in the House election. Treating these estimated quantities as known then

[18] The two-party House vote exceeded the two-party presidential vote in 44 districts in 1988.

[19] Clearly, in the case of Vermont eliminating nonmajor-party House candidates distorts the result, since the Independent candidate garnered a full 37% of the vote while the Democratic candidate finished a distant third. Vermont is quite unusual in this regard, as significant candidates running for neither major party are rare in recent American elections. In fact, Burden and Kimball inadvertently juxtaposed the Independent and the Democrat in the Vermont case, so Table 4 does not describe an observation in their analysis. There are 435 House districts, but Seven Louisiana districts had no contests in November 1988, nine other races were uncontested and produced no vote count, and 65 more were missing a candidate from one of the major parties. Of the 354 remaining districts, 154 saw *some* votes won by nonmajor-party candidates.

TABLE 4 A 2 × 3 Simplification of the Vermont Presidential-House Vote-Splitting Problem

	House of Representatives Vote			
Presidential Vote Choice	**Republican**	**Democrat**	**"None"**	**Total**
Bush (R)	v_{RR}	v_{RD}	v_{Rn}	124331
Dukakis (D)	v_{DR}	v_{DD}	v_{Dn}	115775
Number of Voters	98937	45330	95839	240106

reduces the 2 × 3 problem to a 2 × 2 table where the cell entries are now the quantities of interest, rates of straight and split-ticket voting.

Burden and Kimball's analysis of split-ticket voting in 1988 thus proceeded through multiple stages: (1) they estimated abstention (with EI); (2) they estimated ticket-splitting rates, conditional on the abstention estimates (again with EI); and, (3) they modeled these district-level estimates of split-ticket voting as a function of candidate, institutional, and constituency traits (with OLS). Specification issues arise at every stage of the analysis, and, naturally, the accuracy and validity of each stage depend strongly on the accuracy and validity of the preceding stages. Since the errors from each of the stages compound, the final results are highly prone to indeterminacy. Burden and Kimball make no attempt to incorporate uncertainty from any previous stage of their analysis into the proceeding stages. Instead, at each estimation stage, they begin with a "clean slate," assuming that estimation from previous stages is without error. Even if the estimation at each stage were valid, their multistage estimation procedure poses serious problems.[20]

Each stage of their estimation beginning with the conceptualization of the problem, however, has difficulties. We do not examine the last stage of their estimation closely, but note that Herron and Shotts (2003, 2004) and McCue (2001) have scrutinized the validity of using point estimates generated by EI as dependent variables in a second-stage linear regression, the exact process by which Burden and Kimball arrived at their final estimation. The analysis by Herron and Shotts shows that this process may yield inconsistent and attenuated estimates, but worse, these estimates may suffer from sign reversal and augmentation bias. Clearly, these problems seriously affect the ability to make valid or accurate inferences. Remarkably, Herron and Shotts arrived at these conclusions while assuming that all of the assumptions of EI hold.

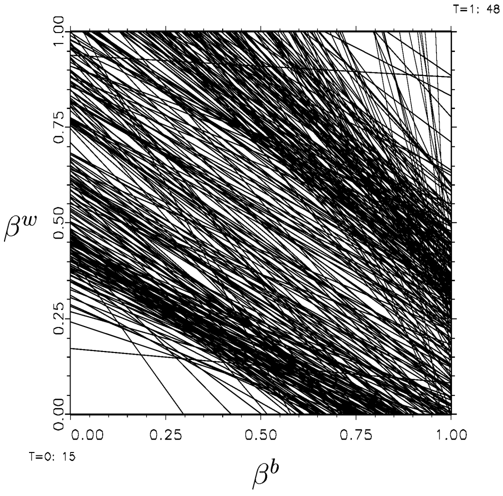
We heed the admonitions of this analysis, but focus on the earlier stages of the Burden and Kimball analysis. Indeed, even if the final stage of their analysis were valid, the preceding stages cast overwhelming doubt in and of themselves. *Every* stage of their analysis is plagued with problems. For conciseness, hereafter, we focus primarily on their stage-two analysis, wherein they produce estimates of district-level split-ticket voting rates. Although the first-stage analysis is less substantively interesting, the problems with EI at the second stage clearly apply to the first-stage analysis as well.

Given the complexity of the American ballot, a full analysis of ticket-splitting is a huge job. For the sake of tractability, Burden and Kimball make some small simplifications (disregarding minor party candidates), some bigger simplifications (disregarding abstention from the presidential contest), some substantial simplifications (examining only two contests at a time), and a *very* strong and unambiguously incorrect assumption that estimation errors produced at each stage of their analysis could be safely ignored thereafter. While these choices made the problem manageable, they also greatly limit the applicability and validity of the eventual substantive conclusions about who splits tickets and why.

Bearing these limitations in mind, do the data on congressional and presidential voting in 1988 nonetheless reveal interesting or novel information about ticket splitting? Following King's own advice, one should begin an ecological inference analysis by assessing how much information is deterministically available in the aggregate data (King 1997, 277–91). Although the authors do not report any diagnostics on the data, and despite the inherent indeterminacy previously discussed, it is a useful first step of analysis, to which we now turn our attention.

[20] How to model the uncertainty as it propagates, across stages, is a critical, but difficult issue. It is not even clear what statistical literature to search for help, since the existence of multiple stages here is motivated not by theory, but by practical computational concerns. One option might be some form of bootstrapping. In the problem at hand, problems with EI's district-level point estimates and their inappropriateness as dependent variables in OLS models overshadow the knotty problem of modelling the compiling of well-behaved uncertainty.

FIGURE 6 Tomography Plot for Burden-Kimball House Vote-Splitting Data



This tomography plot is very dense and uninformative, far more similar to the plot in Figure 2 than to the plot in Figure 1.

Assessing Initial Diagnostics

Consider Figure 6, which displays a tomography plot from the second stage of Burden and Kimball's analysis of the House data set. Recall that for this stage of their analysis, β^b and β^w represent proportion of the Dukakis vote and proportion of the Bush vote (respectively) cast for the Democratic House candidate. This plot resembles the one in Figure 2 in that both are very uninformative. (In fact, the lines in Figure 2 are a random draw of the lines in Figure 6.) The only difference, then, is that the Burden-Kimball tomography plot has more seemingly unrelated lines than the uninformative tomography plot in Figure 2. Again, the bounds are too wide to imply any sort of substantive conclusion. The bounds on β^b are [0.28, 0.91]. The bounds on β^w are [0.24, 0.75]. Even in this initial stage of assessing the information inherent in the data, it would appear that this split-ticket voting data set does not contain much information about the parameters of interest: the bounds are not much narrower than [0, 1], and no general area of intersection is evident. Hence, any inferences made from these data are not likely to be very reliable (King 1997, 185). If the standard errors indicate otherwise, they are likely incorrectly computed.[21] For the Burden and Kimball data then, there is no reason to expect that EI estimates will be reliable. Even if the truncated bivariate normal distribution is a good approximation of the underlying data-generating process, the high variance that characterizes the parameters is likely to render their analysis substantively uninteresting.

Note that we have not yet begun to estimate the parameters of interest. At this initial stage, we are merely assessing how much information is available for the EI estimation procedure. Our initial analyses do not portend success in making correct individual-level inferences based on these aggregate data: the bounds are not informative and no mode is apparent. In some very special situations, when aggregation bias is absent, the method of bounds is truly uninformative yet we are still able to make correct inferences to the individual-level data. So we turn now to the problem of aggregation bias.

Aggregation Bias and Covariate Selection

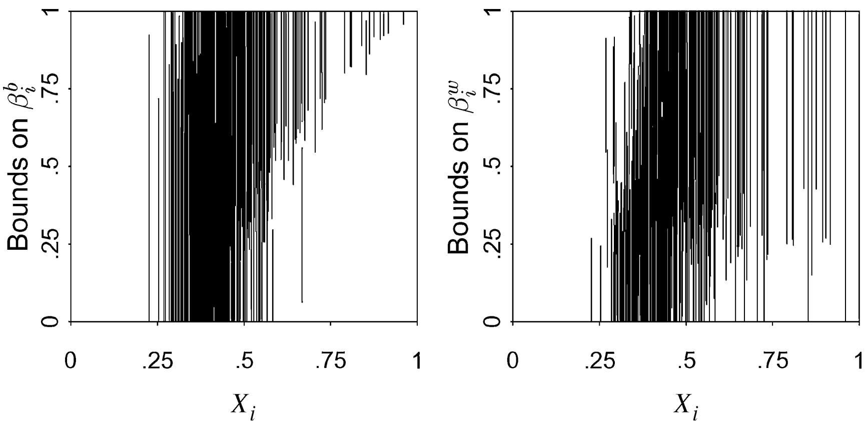
Assessing the degree to which aggregation bias exists is a daunting task, one that is replete with uncertainty. There are, however, some methods that shed some insight into this problem. One method is to examine the aggregation bias diagnostic plot suggested by King (1997, 238). This plot for the data that Burden and Kimball use in their second stage analysis is shown in Figure 7. Aggregation bias exists if there is a relationship between X and β^b or β^w, where, again, X is the proportion of voters who voted for Dukakis. Since β^b and β^w are unknown, we are able to plot only the bounds for these two parameters.[22] We can see from the figure that the vast majority of the bounds cover the entire permissible range from 0 to 1. In addition, the first plot suggests that the aggregation bias for β^b may be severe, since β^b and X appear to be strongly correlated. In other words, there is some evidence that rates of straight-ticket Democratic voting increase with the proportion of the presidential vote won by Dukakis.

A second method for testing whether aggregation bias exists is simply to run the OLS model. If no aggregation

[21] Burden and Kimball puzzlingly state, "Because these are maximum likelihood estimates, King's method also produces standard errors for the two ticket-splitting parameters for each congressional district. In this case it yields fairly precise estimates of ticket-splitting" (1998, 536) They base their assessment of precision on the small standard errors. Again, though, this claim overlooks the fact that their analysis is multistage, and so the errors are compounded. Moreover, it is unclear how much faith can be placed on standard errors when the number of parameters grows with the number of observations (Neyman and Scott 1948). King (1997) does not rigorously analyze the statistical properties of his estimator and so provides no guidance.

[22] And, in this instance, these are not genuine bounds because they incorporate the error from a previous stage of the analysis.

FIGURE 7 Aggregation Bias Diagnostic



Aggregation bias exists if X and β are correlated. The only deterministic information available for β is contained in the bounds. Each line in these plots indicates the range on the bounds for a district. These bounds are generally wide. To the extent that there is any pattern, it indicates a correlation between X and β.

bias exists, the assumptions of OLS are met, and so OLS will yield consistent and unbiased estimates. The OLS model for the Burden and Kimball data yields

$$T = 0.0331 + 1.075X, \tag{3}$$

where X is the proportion of voters who voted for Dukakis, and T is the proportion of the House vote that went to the Democratic candidate. Clearly, OLS does not yield the correct solution. The model predicts Democratic House candidates' vote shares of 3.3% and 110.8% for districts giving 0% and 100% of their vote to Dukakis, respectively. Equivalently, it estimates that 3.3% of Bush voters supported Democratic House candidates and that −10.8% of Dukakis voters supported Republican House candidates. This latter estimate, being logically impossible, alerts us that the assumptions of OLS are violated by these data. Producing out-of-bounds estimates is thus a very useful feature of the linear probability model. While out-of-bounds estimates are clear signals of a misspecified model, the converse of this statement is false: estimates which are within the bounds do not signify a correctly specified model. And since EI *always* produces parameter estimates which are within the [0, 1] bounds, it has no such diagnostic value for assessing whether the specification and/or model assumptions are correct.

Since OLS produces correct estimates if no aggregation bias exists in the data set, one can conclude from equation (3) that there is a high probability of aggregation bias in the data set.[23] Given that EI is not robust to violations of the aggregation bias assumption, and we now have a prior that aggregation bias exists in the data set, the EI estimates are immediately suspect. It is possible that EI will provide reasonable estimates despite the presence of aggregation bias. However, this result would be the exception, not the rule, since EI is a biased and inconsistent estimator in the presence of aggregation bias (Cho 1998; King 1997). The exception might occur when the bounds are informative. Nonetheless, it is clear from Figure 6 that the bounds are far from informative in this 1988 election data set—the vast majority of the bounds span the entire range of possibilities.

One method for mitigating the effects of aggregation bias is to include covariates in the model (King 1997, 288). If these covariates control aggregation bias by accounting for the correlation between the parameters and the regressors, then the model will produce the correct estimates. Burden and Kimball included one covariate in their second-stage model specification, a dummy variable that indicates whether or not a district is located in the South. They included this variable in the belief that individuals who live in the South are unlike individuals who do not live in the South when it comes to decisions

[23] Checking the results from Goodman's regression line is a diagnostic suggested by King: "If Goodman's regression line does not cross both the left and the right vertical axes within the [0, 1] interval, there is a high probability of aggregation bias. If the line does cross both axes within the interval, we have less evidence of whether aggregation bias exists" (King 1997, 282).

about ticket splitting. They offered no indirect evidence (e.g., survey data) to support this contention. Regardless of whether their intuitions about the South are correct or not, including this covariate does not affect the estimates. Their justification for its inclusion was "to account for possible aggregation bias and to improve the estimates" (1998, 536). However, since the two models produce indistinguishable estimates, there is no reason to believe that a South dummy has any desirable effect in mitigating the aggregation bias. If the specification with no covariates is ill-advised, so too is the specification with only the variable "South."

Burden and Kimball were correct that there is a need to alleviate the aggregation bias in their data set and that incorporating the correct covariates would achieve this end. The problem they encountered is that EI does not provide a test for whether one specification is better than another specification. EI users thus find themselves in a truly problematic situation: they cannot determine which specification is correct, but different specifications can produce very different and irreconcilable results. Indeed, in this way, making ecological inferences is no different than more traditional estimation where we have long known that model specification has important consequences for inference. The ecological inference context takes the challenges inherent in any statistical estimation and compounds it with the problems posed by aggregation.

Consider Table 5, which shows the results from different model specifications (the covariates are from the set Burden and Kimball use in their (third-stage) OLS analysis of their EI-estimated split-ticket voting levels). Most of these covariates are associated with some prior theory or result about ticket splitting. Since these are district and candidate attributes, not aggregates of individual voter traits, they are entering the model at the "right" level. (In the next section, we discuss the most important independent variable in Burden and Kimball's analysis, which is, by contrast, partly an aggregate of individual traits and, thus, subject to distortion by aggregation.) But which ones belong in the model? Burden and Kimball incorporated all of them except the South dummy as independent variables in their OLS stage rather than the EI stages not for any theoretical reason, but because EI treats covariates as incidental, and produces no coefficient estimates for them. Of course, the EI model lacking covariates and the OLS follow-up are mutually inconsistent. And "putting off" adding covariates until the OLS stage does not ameliorate the serious problem of aggregation bias in the aggregate analysis. Is there any reason to believe that any of these covariates alleviate the aggregation bias?

TABLE 5 The Effect of Different Covariates

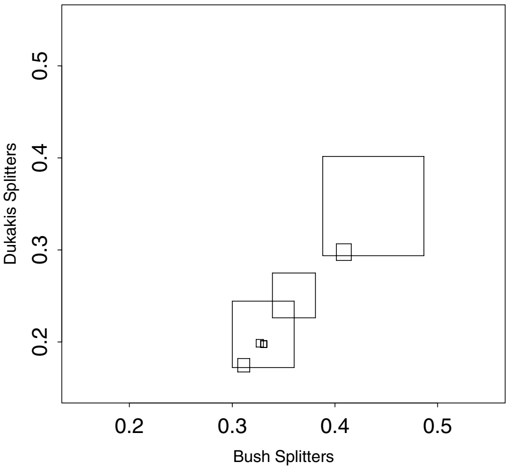
	Bush Splitters	Dukakis Splitters
No Covariates	0.3306	0.1982
	(0.0058)	(0.0074)
South*	0.3310	0.1980
	(0.0060)	(0.0070)
NOMINATE Score^{bw}	0.4376	0.3477
	(0.0986)	(0.1075)
NOMINATE Score^{w}	0.4088	0.2977
	(0.0146)	(0.0180)
Experienced Challenger	0.3602	0.2508
and Money	(0.0421)	(0.0488)
Experienced Challenger	0.3303	0.2085
	(0.0601)	(0.0719)
Democratic Incumbent^{b}	0.3113	0.1749
Republican Incumbent^{w}	(0.0115)	(0.0143)
Ballot^{bw}	0.3269	0.1988
	(0.0068)	(0.0084)
NOMINATE Score^{b}	EI could not estimate	
Experienced Challenger, Democratic and Republican Incumbent	EI could not estimate	

A superscript^{b} indicates the covariate was used for β^{b}.
A superscript^{w} indicates the covariate was used for β^{w}.
*Burden and Kimball's specification.
Standard errors in parentheses

To begin with, in these different specifications, the estimated percentages of ticket splitters vary widely. The values for the standard errors are large for some specifications and extremely small for other specifications. This erratic performance is illustrated in Figure 8. Each rectangle is centered at the point estimate and extends one standard error in each direction. The rectangles would overlap if the models were consistent, but they do not. Even after accounting for the standard error, few of the point estimates are in agreement. Substantively, this is a problem because the alternative specifications imply different types of voting behavior. In addition, the computer program, citing various errors, was not able to compute estimates for certain other specifications. To settle on the best available model of the split-ticket vote, one must somehow choose from among these different specifications. Finding a proper specification is always a major step, but in the aggregate-data context, there are enormous barriers (see Achen and Shively 1995, ch. 4, for an extensive discussion; see also Erbring 1990, 264–65, and Haitovsky 1973).

The advice that King offers is that one should include covariates that can "be justified with specific reference

FIGURE 8 The Effect of Different Covariates



Each rectangle represents the results of a particular specification of the EI model. The rectangles are centered on the point estimate and are one standard error wide and high. If the separate estimates were consistent with one another, the rectangles would all overlap.

to prior substantive knowledge about a problem" (King 1997, 173). He provides no empirical test for choosing covariates, but only this admonition to exercise one's belief about what may be true. This would be unproblematic if different researchers always reached common substantive conclusions after imposing their own beliefs on the model specification. As this congruence virtually never occurs, however, it is obvious that a formal method is needed to determine which covariates are likely to belong in a properly specified model. After all, "including the wrong variables does not help with aggregation bias" (King 1997, 173).

Although the problem of identifying proper covariates with formal tests is not solved, and may not even admit a "solution" in the sense of a universally optimal test, there are some starts on this problem. For instance, Tam (1997) notes that the statistical literature on changepoints and parameter constancy addresses an analogous problem and so is very promising as a source for guidance on how to pick covariates in the aggregate data context. The reason to introduce covariates, after all, is because the parameters of interest are not constant throughout the data set. Hence, a useful empirical test should discriminate between covariates that do divide the sample into subgroups in which parameters are nearly constant and those that do not (Cho 2001). In terms of the TBVN distribution that the EI model incorporates, adding covariates into the model would condition the parameters and allow much more flexibility. We may be interested in testing, for instance, whether people who split their tickets are distinguishable by education level from those who vote straight tickets. If so, we should not necessarily try to fit a TBVN distribution with a single mode and set of variance parameters.

In a general changepoint problem, a random process generates independent observations indexed by some nonrandom factor, often, but not exclusively, time. One may wish to test whether a change occurred in the random process by searching over partitions that divide the data into subsets appearing to have different distribution functions. Again, the subsets can be sorted chronologically, or can be generated from an ordering of some other measured property. The literature is large and diverse: some tests assume that the number of changepoints is unknown, while others assume a fixed number of changepoints; some fix the variance of the distribution, while others estimate the variance as a parameter; some assume that the different distributions take similar forms, while others allow more flexibility in this regard. Tests vary in nature as well: some are Bayesian (e.g., Carlin, Gelfand, and Smith 1992; Schulze 1982; Smith 1975), some are parametric (e.g., Andrews, Lee, and Ploberger 1996; Ritov 1990), some are nonparametric (e.g., Carlstein 1986; Wolfe and Schechtman 1984), and some are related to time series analysis (e.g.; Brown, Durbin, and Evans 1975). In short, there are diverse means by which one can draw inferences about changepoints and constancy, or lack thereof, of parameters. For present purposes, what is important is that the general object in this literature is to find a means for partitioning data sets into subsets within which there is some degree of parameter constancy. Since this is precisely the goal for the researcher choosing covariates in an ecological inference problem, the application of changepoint tests to aggregate data problems seems extremely promising.

Cho (2001) introduces one formal covariate-selection test adapted from time-series analogs. There is not likely to be *one* covariate-selection test that is optimal for all aggregate data problems, but employing an empirical test is clearly preferable to imposing subjective beliefs. It is important that aggregate data analysts have some standard by which to judge whether one specification is superior to another. Further development of well-specified statistical tests for covariate selection should be the priority for aggregate data research.

Burden and Kimball were in need of just such a test, since their data exhibited aggregation bias. Lacking any means by which to compare covariates that might alleviate the problem, they settled on one covariate chosen on qualitative grounds. Unfortunately, this covariate did not perform the necessary function of removing aggregation bias, and their analysis suffered accordingly.

Microtheory: Intention and Ticket Splitting

Burden and Kimball contend that their research makes two distinct contributions to the study of ticket splitting. First, as pioneers in applying King's EI methods, they purport to provide the first accurate estimates of the extent of ticket splitting. Second, their analysis of splitting (as estimated by EI) reveals that it is primarily an unintentional rather than intentional activity. Americans simultaneously support different parties at a given moment not because they prefer to see power balanced or shared, but because strategic choices by candidates and parties induce splitting. We have already demonstrated that the first of these innovations is more apparent than real—EI certainly does not produce new levels of accuracy in estimating ticket-splitting behavior. Their data were neither informative nor immune to aggregation bias. We now discuss the second main reason to doubt the generality and veracity of their conclusions, namely that their test was not informed by serious consideration of microlevel theories.

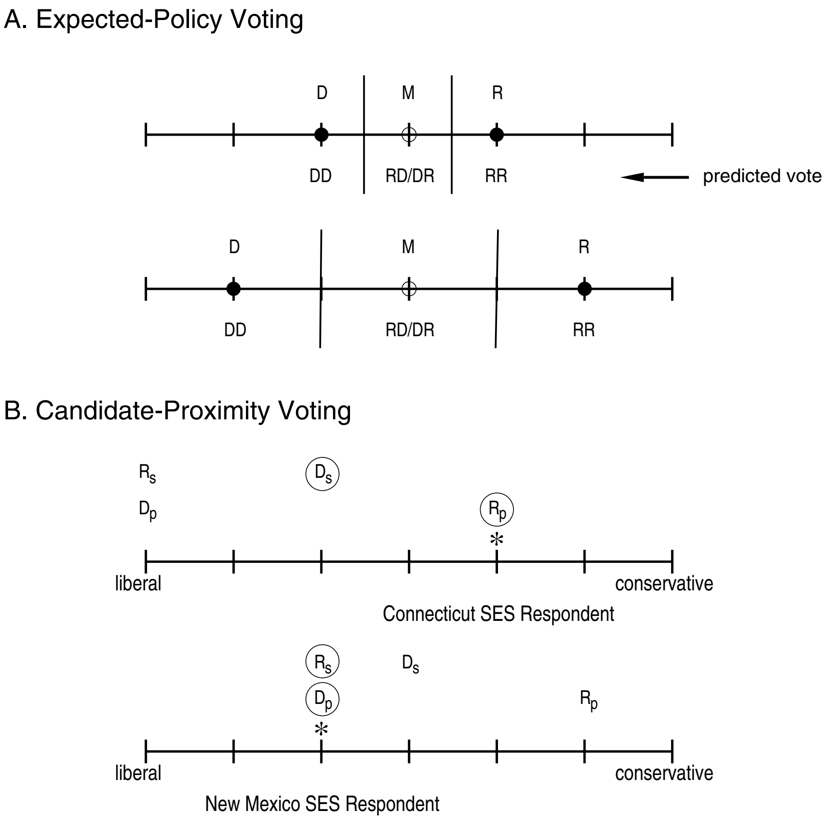
The term "intentional" could be ambiguous in this context, but the authors clarify that their interest lies in assigning primary responsibility for ticket-splitting to either candidates or voters (Burden and Kimball 1998, 533). If levels of tickets-splitting seem to respond to candidate traits such as incumbency, spending differentials, or candidate experience, they propose, it is not the case that the masses deliberately divide their support between parties. Thus, they conclude, the candidates, not the voters, move first. It is, of course, already very well known that contemporary American elections feature a substantial incumbency advantage. To verify that some ticket splitting seems to originate in incumbents' skills at drawing cross-party support, however, is not to rule out that voters are quite consciously spreading support across parties or ideologies. Burden and Kimball did not test whether incumbents are helped or hindered in drawing nonparty-based support by the expected fates of their parties' presidential candidates. In that respect, they do not give "intentional" ticket-splitting much chance to surface.

The critical result for their claims about balancing and intent, ultimately, is an insignificant coefficient on the variable they label "ideological distance." They operationalize this variable as the mean distance on a seven-point ideology scale between Bush (R_p) and the Democratic Senate candidate (D_s), as assigned by a state's Senate Election Study (SES) respondents. There are some problems with this construction. Aggregation to state means is noisy: in most states, standard deviations of R_p and D_s span about a quarter of the entire interval. More importantly, the idea that greater spread between these two candidates might yield more ticket splitting relies on some strong, unstated assumptions about intraparty homogeneity, voter distributions, and the origin of vote splitting.

Figure 9A illustrates the logic whereby spatial party differentials might lead to vote splitting. If both Democrats (e.g., D_s and D_p, where *s* and *p* denote "senate" and "presidential" candidates) are located at *D*, both Republicans at *R*, and if expected policy outcomes for unified government are, thus, *D* and *R*, but for divided government are some weighted average of *D* and *R*, say *M*, then standard proximity theory identifies cutpoints defining zones in which voters should prefer to vote straight tickets (*DD* or *RR*) or split tickets (*DR* or *RD*). Then, as $|D_s - R_p|$ grows, the central region containing split-ticket voters grows. Burden and Kimball's "ideological distance" variable is thus constructed on three assumptions: first, that voters react to expected policy outcomes, not candidates per se; second, that the two Democrats and two Republicans in question are ideologically very similar, if not identical; and, third, that a substantial portion of the electorate resides in the center of the ideological spectrum, so that enlargement of the middle region does result in more split-ticket voting occurring.

Note, then, that the logic of their test fails if voters do not perceive the two Republicans and two Democrats to be ideological twins or if the district's electorate is bipolar. On the first point, consider Figure 9B, and suppose that Connecticut has a symmetrical (e.g., uniform) voter distribution. As R_p moves right, $|D_s - R_p|$ increases. Under the same assumptions about voting as just applied to Figure 9A, though, the amount of split ticket voting should *decrease* with this increase in $|D_s - R_p|$, since the ticket splitters are now those in the outer regions, given this particular arrangement of candidates. This is the *exact opposite* effect from that shown by Figure 9A and assumed to apply everywhere by Burden and Kimball. And, although they point out in their footnote 11 (1998, 539) that the SES set uniquely provides the necessary data to test a balancing thesis since it includes placements of respondents and *all four candidates* on a common scale, only half of this information is actually incorporated in their construction of "ideological distance." Either large ideological variation between different party nominees or a noncentrally distributed electorate thwarts interpretation of their regression results.

FIGURE 9 Alternative Spatial Theories of Split-Ticket Voting


A. Expected-Policy Voting
D
M
R
DD
RD/DR
RR
predicted vote
D
M
R
DD
RD/DR
RR
B. Candidate-Proximity Voting
Rs
Ds
Dp
Rp
*
liberal
conservative
Connecticut SES Respondent
Rs
Ds
Dp
Rp
*
liberal
conservative
New Mexico SES Respondent

Predicted and actual votes cast are circled in part B. For these two respondents, actual votes are identical to votes as predicted by expected-policy and candidate-proximity theories.

There is, moreover, a plausible variety of intentional vote splitting not captured by their logic. Voters who select candidates only according to ideological proximity, without making projections about policy outcomes that will result from the various permutations of candidate victories, can intentionally split tickets if they perceive there to be large differences in the positions of candidates from the same party. Figure 9B shows two actual SES respondents (marked with asterisks) whose split-ticket votes exactly match the predictions of simple candidate-proximity theory. Note that $|D_s - R_p|$ is identical in the two cases, even though one is a vote of R_pD_s and the other a vote of D_pR_s. For both of these respondents, in fact, reported votes are consistent with either expected-policy voting or candidate-proximity voting. This observational equivalence also undercuts strong claims about voter intentions.

More generally, spatial theories of voting are many and varied, and even if one posits that voters choose according to policy outcomes, the proper econometric specification to test for "intent" will depend critically on the underlying formal model. Merrill and Grofman's recent work (1999) unifying directional and proximity models is one excellent blueprint in this regard, since they carefully construct a hybrid model in which both rivals are nested, and allow data to adjudicate between them, or to select a mixture.

Even more pertinent to the issue at hand are two recent articles about how American voters do or do not link their votes in search of moderate policy. In his analysis of individual-level voting data from the NES, Mebane reaches a conclusion very different from Burden and Kimball, that "policy-related balancing has often been an important determinant of election outcomes." (2000, 51) Mebane and Sekhon (2002) extend the logic of that article to midterm voting and find further evidence for not only moderation, but coordination. In both cases, a virtue in these articles is that coordinating and noncoordinating models are estimated in tandem, so that competing models can be compared formally. In each case, coordination and moderating are formally and explicitly defined and distinguished from related phenomena such as economic retrospective voting and incumbency advantage. Mebane notes that the model does not achieve great success at identifying ticket-splitters (2000, note 26, 51), and this work is not necessarily the last word on the topic. But the care with which the terms of the theoretical model are operationalized is exemplary, and the conclusions are appropriately qualified, particularly concerning underlying assumptions about the effects of institutional context, and how these relate to unrealistic assumptions about voter-level mechanisms. Burden and Kimball's analysis, by contrast, is not nearly flexible or general enough to support their strong conclusion that voters do not consciously choose to split tickets.

Conclusion

Our hope is that our discussion of ecological inference and its inherent uncertainty highlights the numerous reasons why a researcher must exercise great caution when analyzing aggregate data. With EI, in particular, the purported advances are coupled with much greater computational complexity and a large number of new assumptions. Since EI generally does not outperform OLS (Cho 1998), it is difficult to justify the additional overhead. EI does supply district-level estimates, but these estimates do not possess desirable statistical properties (Herron and Shotts 2003, 2004). There are instances when one needs to make ecological inferences, and so one will choose to use an ecological inference model such as EI. In these cases, the researcher must be fully aware of the numerous pitfalls that may ensue.

For example, Burden and Kimball claim as their principal achievement to have developed the first-ever accurate district-level estimates of vote-splitting. However, their faith in the multi-stage aggregate data analysis is misplaced, and there are a multitude of reasons to doubt the accuracy of their findings. Furthermore, their search for "intention" in vote-splitting is more accurately a test for a particular kind of balancing behavior, under key assumptions about ideological homogeneity within parties and distributions of district electorates. That analysis is not general enough to support strong conclusions about voting behavior. It is unreasonable to declare new-found knowledge when the novel findings depend critically on very strong and unverifiable assumptions about the underlying individual-level data. Split-ticket voting behavior remains a fascinating topic, and it also remains a topic plagued by severe data-analysis barriers.

Caution can never be thrown to the wind when an analysis proceeds through multiple stages of analysis, especially when multiple stages of ecological inference are involved. Ultimately, there is no escaping indeterminacy in cross-level inference. The problem is ill-posed and so not amenable to unique "solutions" as such. Here we have emphasized that those who must proceed with ecological inference just the same ought to know their data well; be aware that even ostensibly informative data can be misleading; be on guard against aggregation bias, and endeavor to model it when it occurs; and be as explicit as possible about the logic connecting the micro- and macrolevels. But a final caveat is that one can still go astray even having exercised care in all of these manners. Hence, except in highly unusual circumstances, aggregate data analysis intended to yield insight into micro behavior always calls for cautious and guarded interpretation.

Appendix

Software Issues

Although we did not discuss computer programming and model implementation here, we note that the question of whether King's software reliably implements his statistical model has been raised by others, in passing (Freedman et al. 1998, 1520; 1999, 356) and in great detail (Altman and McDonald 2001). The Freedman et al. review of King's book notes that their independent coding of the EI model yielded different results from those output by King's software. Altman and McDonald conclude that the changes between different versions of the software "can produce differences in estimates large enough to affect the substantive conclusions made on the basis of an EI analysis" (2001, 1). In addition, King has revised the program many times and has identified numerous changes and bug fixes in the "What's New?" documentation that accompanies the software. Burden and Kimball obtained their results using version 1.21 of *EzI*, one of the earliest versions of the program, that predates many bug fixes.

For instance, the "What's New" documentation notes: "9/15/96 fixed a small buglet for unanimous precincts"and "9/25/96 fixed a bug that affected homogenous precincts rarely." (These fixes appear to have been made in response to a paper by Rivers and Tam (1996) that identified a mistake in his derivation of the likelihood.) King later realized that his program produced significant differences according to whether it was run on a Windows machine or a Unix machine ("1/13/97 Unified DOS/Unix version to cover differences in Gauss across platforms"). Despite all these updates, Altman and McDonald report that there are still differences in its operation across platforms (2001, 11). In short, King has made (and continues to make) many corrections and changes to his software, but it is not clear how those bugs he has fixed might have affected previous results. Some of these bug fixes involve procedures that Burden and Kimball employed, such as EI2 ("3/31/98 Fixed bug in EI2"). Moreover, there is an ongoing debate about the reliability of the GAUSS programming language on which King's program depends. Inaccuracies can appear when performing basic statistical computations such as linear and nonlinear regressions, simulations, and *t*-values with GAUSS. In some evaluations, the number of accurate digits produced by GAUSS is zero (McCullough and Vinod 1999; Vinod 2000). Hence, users of *EzI* may be subject to multiple layers of programming errors. While virtually all data analysis relies on software, and thus potentially inherits problems of implementation, the cause for concern is clearly greater with *EzI*.

References

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

Alesina, Alberto, and Howard Rosenthal. 1995. *Partisan Politics, Divided Government, and the Economy*. Cambridge: Cambridge University Press.

Altman, Micah, and Michael McDonald. 2001. "Ensuring Numerical Stability in Ecological Inference." Working paper. Harvard University.

Andrews, Donald W. K., Inpyo Lee, and Werner Ploberger. 1996. "Optimal Changepoint Tests for Normal Linear Regression." *Journal of Econometrics* 70(1):9–38.

Anselin, Luc, and Wendy K. Tam Cho. 2002. "Spatial Effects and Ecological Inference." *Political Analysis* 10(3):276–97.

Brown, R. L., J. Durbin, and J. M. Evans. 1975. "Techniques for Testing the Constancy of Regression Relationships Over Time." *Journal of the Royal Statistical Society, Series B* 37(2):149–92.

Burden, Barry C., and David C. Kimball. 1998. "A New Approach to the Study of Ticket Splitting." *American Political Science Review* 92(3):533–44.

Carlin, Bradley P., Alan E. Gelfand, and Adrian F. M. Smith. 1992. "Hierarchical Bayesian Analysis of Changepoint Problems." *Applied Statistics* 41(2):389–405.

Carlstein, E. 1988. "Nonparametric Change-Point Estimation." *Annals of Statistics* 16(1):188–97.

Cho, Wendy K. Tam. 1998. "Iff the Assumption Fits. . . : A Comment on the King Ecological Inference Solution." *Political Analysis* 7:143–63.

Cho, Wendy K. Tam. 2001. "Latent Groups and Cross-Level Inferences." *Electoral Studies* 20(2):243–63.

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):237–64.

Duncan, Otis Dudley, and Beverly Davis. 1953. "An Alternative to Ecological Correlation." *American Sociological Review* 18(6):665–66.

Erbring, Lutz. 1990. "Individuals Writ Large: An Epilogue on the 'Ecological Fallacy."' *Political Analysis* 1:235–69.

Fiorina, Morris P. 1996. *Divided Government*, 2^{d} ed. Needham Heights, MA: Allyn and Bacon.

Freedman, D. A., S. P. Klein, M. Ostland, and M. R. Roberts. 1998. "A Solution to the Ecological Inference Problem." *Journal of the American Statistical Association* 93(444):1518–22.

Freedman, D. A., M. Ostland, M. R. Roberts, and S. P. Klein. 1999. "Response to King's Comment." *Journal of the American Statistical Association* 94(445):355–57.

Gehlke, C. E., and Katherine Biehl. 1934. "Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material." *Journal of the American Statistical Association* 29(185):169–70.

Goodman, Leo A. 1953. "Ecological Regressions and Behavior of Individuals." *American Sociological Review* 18(6):663–64.

Goodman, Leo A. 1959. "Some Alternatives to Ecological Correlation." *American Journal of Sociology* 64(6):610–25.

Haitovsky, Yoel. 1973. *Regression Estimation from Grouped Observations*. New York: Hafner.

Herron, Michael C., and Kenneth W. Shotts. 2003. "Using Ecological Inference Point Estimates as Dependent Variables in Second-Stage Linear Regressions." *Political Analysis* 11(1):44–64.

Herron, Michael C., and Kenneth W. Shotts. 2004. "Logical Inconsistency in EI-based Second Stage Regressions." *American Journal of Political Science* 48(1):171–82.

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

King, Gary. 1998. "EI: A Program for Ecological Inference." Version 1.18, February 20. Harvard University.

King, Gary. 1999. "The Future of Ecological Inference Research: A Comment on Freedman et al." Letter to the Editor of the *Journal of the American Statistical Association* 94(445):352–55.

Kramer, Gerald H. 1983. "The Ecological Fallacy Revisited: Aggregate- versus Individual-level Findings on Economics and Elections, and Sociotropic Voting." *American Political Science Review* 77(1):92–111.

McCue, Kenneth F. 2001. "The Actual Value of EI-R Coefficients." Working Paper. CalTech.

McCullough, B. D., and H. D. Vinod, 1999. "The Numerical Reliability of Econometric Software." *Journal of Economic Literature* 37(2):633–65.

Mebane, Walter R., Jr. 2000. "Coordination, Moderation, and Institutional Balancing in American Presidential and House Elections." *American Political Science Review* 94(1):37–57.

Mebane, Walter R., Jr., and Jasjeet S. Sekhon. 2002. "Coordination and Policy Moderation at Midterm." *American Political Science Review* 96(1):141–57.

Merrill, Samuel, III, and Bernard Grofman. 1999. *A Unified Theory of Voting: Directional and Proximity Spatial Models*. Cambridge: Cambridge University Press.

Neyman, J., and Elizabeth L. Scott. 1948. "Consistent Estimates Based on Partially Consistent Observations." *Econometrica* 16(1):1–32.

Ritov, Y. 1990. "Asymptotic Efficient Estimation of the Change Point with Unknown Distributions." *Annals of Statistics* 18(4):1829–39.

Rivers, Douglas, and Wendy K. Tam. 1996. "Estimation of Random Coefficient Models." Working paper. Stanford University.

Robinson, W. S., 1950. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review* 15(3):351–57.

Schulze, U. 1982. "Estimation in Segmented Regression: Known Number of Regimes." *Mathematische Operationsforschung und Statistik, Series Statistics* 13: 295–316.

Smith, A. F. M. 1975. "A Bayesian Approach to Inference About a Change-point in a Sequence of Random Variables." *Biometrika* 62(2):407–16.

Tam, Wendy K. 1997. "Structural Shifts and Deterministic Regime Switching in Aggregate Data Analysis." Master's Essay. Department of Statistics. University of California at Berkeley.

Vinod, H. D. 2000. "Review of GAUSS for Windows, Including Its Numerical Accuracy." *Journal of Applied Econometrics* 15(2):211–20.

Wolfe, Douglas A., and Edna Schechtman. 1984. "Nonparametric Statistical Procedures for the Changepoint Problem." *Journal of Statistical Planning and Inference* 9(3):389–96.