Iff the Assumption Fits . . . : A Comment on the King Ecological Inference Solution

Wendy K. Tam Cho

Abstract

I examine a recently proposed solution to the ecological inference problem (King 1997). It is asserted that the proposed model is able to reconstruct individual-level behavior from aggregate data. I discuss in detail both the benefits and limitations of this model. The assumptions of the basic model are often inappropriate for instances of aggregate data. The extended version of the model is able to correct for some of these limitations. However, it is difficult in most cases to apply the extended model properly.

Introduction

For a wide variety of questions, especially those which involve historical research or volatile issues such as race, aggregate data supply one of the only sources of reliable data. However, making inferences about micro-level units when the only available data are aggregated above the micro-level unit in question is extremely difficult. Consider the example of aggregate data analysis in investigating the gender gap by estimating presidential voting behavior among women. For each precinct, the

Thanks to Bruce Cain, Douglas Rivers, Gary King, Walter Mebane, Henry Brady, Chris Achen, David Freedman, Brian Gaines, Jasjeet Sekhon, Simon Jackman, Michael Herron, Rui de Figueiredo, Jake Bowers, Cara Wong, Mike Alvarez, and Jonathan Nagler for helpful comments. I am also grateful to participants at a CIC Video Methods Seminar and at the University of California, Berkeley, for lively and helpful discussions, and to the National Science Foundation (Grant No. SBR-9806448) for research support.

number of votes received by each presidential candidate is known. In addition, the number of registered female voters and the number of registered male voters is known. However, because of the secret ballot, the manner in which the votes were cast is unknown. In the absence of reliable survey data, one can determine the number of men and women who voted for each candidate only by modeling the situation. There are a variety of assumptions and many statistical methods upon which such a model can be formed.

Whenever a statistical model is employed, one should consider the implications of the model's assumptions. All aggregate data models incorporate assumptions which must be taken under careful consideration. The merits of applying a statistical model to a problem necessarily depend on the consistency of the assumptions with the problem at hand. The King (1997) model (hereafter referred to as "EI") for reconstructing individual behavior from aggregate data is no exception. In this paper, I examine the benefits and limitations of applying EI to instances of aggregate data.

The Basic EI Model

There are basically two "flavors" of EI. For ease, one will be called "basic EI" and the other will be referred to as "extended EI." Basic EI merits its own discussion because it is claimed to be adequate in many situations (King 1997, 284). For this reason, I discuss basic EI first, and then examine the extended model, with particular attention to its ability to compensate for the shortcomings of the basic model.

The Basic Model. EI is a notable advancement to ecological inference in that it incorporates two ideas which are ideally suited for aggregate data, but which have never previously been utilized in tandem. Together, these elements bring a new degree of efficiency to aggregate data analysis. The first of these two ideas is the deterministic method of bounds, first introduced by Duncan and Davis (1953). The method of bounds narrows the range of possible parameter estimates. Since we are normally interested in estimating probabilities or proportions, the range of possible parameter values is immediately restricted to the closed interval $[0, 1]$. Through the method of bounds, we can usually restrict this range even further. However, while the method of bounds can, in theory, provide extremely narrow bounds, it rarely does so in practice. The EI bounds are an obvious extension of the Duncan and Davis bounds and have been known and used previously (Shively 1974; Ansolabehere and Rivers 1997). The additional information incorported in these bounds is still not generally sufficient to make interesting substantive claims.

Since the method of bounds is generally not sufficient in and of

itself, EI incorporates a second probabilistic component, the random coefficient model, largely popularized by Swamy (1971). In particular, EI assumes that the parameters are not constant and that the parameter variation can be described by a truncated bivariate normal distribution. In other words, the assumption is that the parameters "have something in common—that they vary but are at least partly dependent upon one another" (King 1997, 93). The suggestion that a random coefficient model may be better suited than OLS for aggregate data was originally made by Goodman (1959). King's contribution, then, is just the choice of distribution.

The basic EI model incorporates three assumptions (King 1997, 158). First, the parameters are assumed to be distributed according to a truncated bivariate normal distribution. Second, the parameters are assumed to be uncorrelated with the regressors. In other words, "aggregation bias" is not present (King 1997, 55). Lastly, it is assumed that the data do not exhibit any spatial autocorrelation. Since the basic model is not appropriate for every instance of aggregate data (King 1997, 24), it is useful to determine how robust the basic model is to deviations from its assumptions and thus to determine when the basic model is appropriate.

Monte Carlo Simulations. The assumptions of the model can be examined through Monte Carlo simulations. King (1997) performed some of these tests for basic EI. In particular, one Monte Carlo experiment with data inconsistent with the spatial autocorrelation assumption but consistent with the distributional assumption and the assumption of uncorrelated parameters and regressors was performed (King 1997, 168, Table 9.1). Another Monte Carlo simulation included data which were inconsistent with the distributional assumption but consistent with the spatial autocorrelation assumption and the assumption of uncorrelated parameters and regressors (King 1997, 189, Table 9.2). A final Monte Carlo simulation generated data inconsistent with the no aggregation bias assumption but consistent with both the distributional and spatial autocorrelation assumptions (King 1997, 179–182, $\tau = 0$ case). Following are three Monte Carlo simulations which exactly replicate King's setup. In addition, to gain a benchmark for comparison, the results are compared to those obtained using OLS as the aggregate data model.

Consider first the consequences of spatial autocorrelation in aggregate data. The results of a Monte Carlo simulation are reported in Table 1. Each row of the table summarizes 250 simulations drawn from the model with the degree of spatial autocorrelation δ and number of observations p.[1] The data were generated in exactly the same manner as

[1] Due to computational problems resulting from a lack of RAM, the simulations

TABLE 1. Consequences of Spatial Autocorrelation

	OLS			Basic EI		
δ	p	Error	(S.D.)	p	Error	(S.D.)
0	100	.0224	(.0167)	100	.001	(.020)
0	750	.0085	(.0065)	1,000	.000	(.007)
.3	100	.0227	(.0165)	100	.001	(.020)
.3	750	.0082	(.0063)	1,000	.000	(.006)
.7	100	.0221	(.0161)	100	.001	(.022)
.7	750	.0079	(.0060)	1,000	.001	(.006)

described in King (1997, 166). The reported results for basic EI are taken directly from Table 9.1 in King (1997, 168).

While one might expect spatial autocorrelation to be problematic in aggregate analysis, this is clearly not the case if the data are consistent with the other two assumptions. The Monte Carlo evidence implies that spatial autocorrelation, on its own, does not induce bias into either the OLS or EI model. While the error for EI is smaller and approaches zero faster, the OLS results are similarly favorable. Certainly, one would be thrilled with an aggregate data model that performs as well as the OLS model does on these data. Indeed, both of these models are robust against deviations from the spatial autocorrelation assumption when it is the only inconsistent assumption.

A second Monte Carlo experiment examines the consequences of data that are inconsistent with the distributional assumption but exhibit neither aggregation bias nor spatial autocorrelation. These data were generated according to the exact data generating process described in King (1997, 188) for the truncated normal distribution. In order to maintain consistency with the other two assumptions, the parameters were chosen so that truncation is symmetric. Each distribution has two modes which differ but have the same variance/covariance structure.

The results are displayed in Table 2. The numbers reported for basic EI are taken directly from King, Table 9.2, p. 189. The point estimates from the basic EI model seem to be better than the point estimates from the OLS model. However, once we take the standard deviation into account, the estimates are indistinguishable. Both models perform quite admirably when faced with distributional misspecification if the spatial autocorrelation and aggregation bias assumptions hold. Again, robustness to the distributional assumption is clear if all other

with 1,000 observations that were performed for EI could not be replicated. However, simulations with 750 observations were performed and are equally sufficient in establishing the same pattern as n increases.

TABLE 2. Consequences of Distributional Misspecification

		OLS		Basic EI	
Truncation	p	Error	(S.D.)	Error	(S.D.)
Low	100	.0157	(.0150)	.001	(.020)
High	100	.0168	(.0142)	.001	(.011)
Low	25	.0289	(.0254)	.001	(.038)
High	25	.0289	(.0250)	.001	(.024)

assumptions are consistent.

Lastly, data that exhibit aggregation bias but are consistent with the distributional and spatial autocorrelation assumptions were generated. In this simulation, 250 data sets were generated exactly according to the description in King (1997, 161). King describes these data as a "worst case scenario" because, he says, the data have bounds that are minimally informative (1997, 161, 182). I generated the data randomly from the model with parameters $\beta^b = \beta^w = 0.5$, $\sigma_b = 0.4$, $\sigma_w = 0.1$, and $\rho = 0.2$. The results are displayed in Figures 1 and 2. The true parameter values, $\beta^b = \beta^w = 0.5$, are marked in the plots by a vertical line.

Indisputably, these results are orders of magnitude worse than the results of the first two simulations. The density plots in Figure 1 clearly show that the point estimates are far from the true values. Figure 2 plots the error bars. For each simulation, a bar is drawn where the center of the bar is the point estimate. The bar extends one standard error to the left and one standard error to the right. As we can see, the error bars in Figure 2 clearly indicate that, even accounting for the standard errors, the estimates are inaccurate.[2] Moreover, the sense of precision is overstated more by EI than OLS. On average, the EI estimates for β^b are 25 S.E.s from the true value. For β^w, the EI estimates are, on average, -14.7 S.E.s from the true value. Compare these results with the OLS results. On average, in the OLS model, 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. Obviously, the standard errors are erroneously estimated and suggest more precision than actually exists. Although inconsistencies with the distributional assumption and the spatial autocorrelation assumption are not consequential if aggregation bias does not simultaneously exist, this auspicious condition does not hold for the aggregation bias assumption. Even if the data are consistent with the other two assumptions, if the

[2] There are two instances out of 250 simulations where the error bars touch the true parameter values. However, these two instances clearly seem to be anomalies and the result of erroneous calculations by the EzI estimation program.

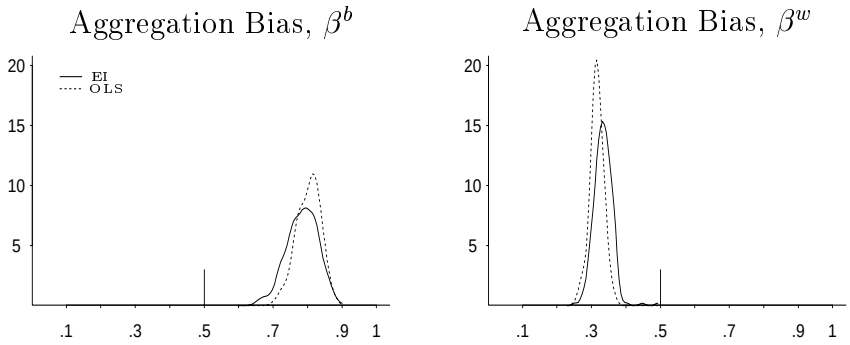

Fig. 1. Density plots from a Monte Carlo simulation with data which are consistent with the distributional and spatial autocorrelation assumptions but inconsistent with the aggregation bias assumption. The true value of the parameter is marked by a small vertical line.

parameters are correlated with the regressors, neither OLS nor EI will yield accurate results. Neither model displays any noticeable robustness to this assumption.

Despite the obvious lack of robustness here, King states that the use of the bounds can make basic EI "robust," "even in the face of massive aggregation bias," because if the bounds are informative, they will "provide a deterministic guarantee on the maximum risk a researcher will have to endure, no matter how massive aggregation bias is" (King 1997, 177, 182). He writes, "Under the model introduced here, 'aggregation bias' in the data does not necessarily generate biased estimates of the quantities of interest" (King 1997, 218). He also provides empirical examples that show that when the bounds are informative, EI can be "robust" in King's sense of the word even though there is aggregation bias. In Chapter 11, the posterior distribution of the state-wide fractions covers the true values very well (King 1997, 222). In Chapters 12 and 13, the posterior distribution covers the true values well for one of the fractions but not the other (King 1997, 231, 240). In King's view, informative bounds are necessary for EI to be "robust," again in his special sense of "robust," when there is aggregation bias.

Despite the reference to "risk," King's conception of robustness has no connection to formal treatments of robustness to assumptions (robust priors) such as have been developed in Bayesian statistical theory (Berger 1985). King's notion of robustness is also unrelated to formu-

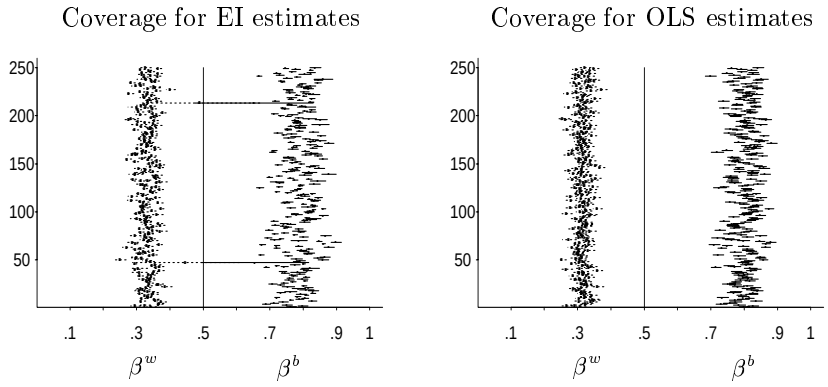

Fig. 2. Error bar plots from a Monte Carlo simulation with data which are consistent with the distributional and spatial autocorrelation assumptions but inconsistent with the aggregation bias assumption. The true parameter values are marked by the long vertical lines. The error bars to the left of the 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.

lations such as those of Box (1953) and Scheffé (1959) which define a robust method as one in which the inferences are not seriously invalidated by the violation of assumptions. Nor is King's definition consistent with the work of Huber (1981), which defines an estimator as robust if it is consistent even when part of the data is contaminated. King does not assert that the use of bounds in basic EI means that basic EI is an unbiased or consistent estimator if there is aggregation bias. Indeed, it is not. If there is aggregation bias, the basic EI estimator is biased, and the discrepancy between the estimates and the true values does not converge in probability to zero as the sample of data points becomes large.

One should note that the data for these experiments are rather artificial. One would not expect to see such patterns in real instances of aggregate data. The three assumptions of the basic EI model are logically distinct, but data will not often be consistent with one assumption while inconsistent with the other two assumptions (King 1997, 159). More likely, the data will be inconsistent with more than one assumption. The Monte Carlo experiments have demonstrated that the crucial assumption concerns aggregation bias. In other words, if the parameters are not correlated with the regressors, aggregate data analysis is not

TABLE 3. Hypothetical Aggregate Data Set for Presidential Vote

		Vote for Clinton			Total Number of	
	Precinct Leaning	Total	From Minorities	From Majority	Minority Voters	Majority Voters
1	Democrat	128	56	72	80	120
2	Republican	72	30	42	60	140
3	Democrat	130	70	60	100	100
4	Republican	74	35	39	70	130
5	Democrat	134	98	36	140	60
6	Republican	80	50	30	100	100
Ecological Regression			Minority Vote: 90%		Majority Vote: 20%	
Basic EI			Minority Vote: 77%		Majority Vote: 30%	
Truth			Minority Vote: 62%		Majority Vote: 43%	

problematic.

Hypothetical Example. A final example of artificial data illustrates in a particularly simple way how correlation between regressors and parameters causes difficulty for ecological inference. Consider the hypothetical data in Table 3. The goal is to determine rates of voting for Clinton among minority voters and majority voters. The true majority support for Clinton is 43 percent while the true minority support is 62 percent. In a contrived example, this information is easily retrieved. Minority support for Clinton in precinct 1 is $56/80 = 70$ percent, and so on. In general, however, even though we can obtain the number of minority voters in each precinct, the number of minorities who voted for Clinton is unknown. Instead, Clinton's support among different groups must be modeled.

The OLS model (also referred to as "Goodman's regression") is

$$(\%\ \text{CLINTON VOTE}) = (1 - \%\ \text{MINORITY})\beta^M + (\%\ \text{MINORITY})\beta^m + e$$

where $e \sim N(0, \sigma^2)$. OLS assumes that the parameters are constant regardless of precinct of residence; i.e., in precinct 1, minority support for Clinton is m percent, and minority support in precincts 2–6 is the same m percent. Since minority support in precinct 1 is 70 percent while minority support in precinct 2 is 50 percent, the assumption of constancy is plainly wrong. Hence, it is not surprising that the OLS model mistakenly reports that 90 percent of the minorities voted for Clinton while 20 percent of the majority voted for Clinton. The problem is that minorities who live in precincts which lean Democratic support Clinton at higher rates than those who reside in more Republican precincts. In other

words, the parameters are correlated with the regressors. As a result of this correlation, the parameter estimates are biased (Ansolabehere and Rivers 1997).

EI accounts for the parameter variation by assuming that while the parameters are not constant, they retain a single common mode that is described by a truncated bivariate normal distribution. In this example, it is clear that this distributional assumption is inappropriate. The distribution is unimodal but the data it purports to describe are bimodal, one mode for the Republican districts and one mode for the Democratic districts. The incorrect distributional assumption simultaneously exists with the correlation between the parameters and the regressors. Hence, also not surprisingly, basic EI produces poor estimates of the true parameters.[3] Basic EI, like OLS, will produce poor results when its assumptions do not fit the data.

Some of the stringent assumptions of basic EI can be modified in extended EI. In the extended EI model, the components of basic EI, bounds and varying parameters, remain the same. However, extended EI allows a user to modify the distributional assumption for the parameter variation by including covariates to describe separate modes.[4] After conditioning on covariates, if the parameters are mean independent of the regressors, aggregate data analysis is straightforward. In this case, the precinct's partisan leaning is the crucial missing covariate. Adding this covariate to the model allows the Democratic precincts to have one mean while the Republican precincts would have a separate mean. This setup defines two subsets of the data where the parameters are constant and therefore clearly not correlated with the regressors. Obviously, if one can identify subsets of the data where the parameters do not vary, the correlation between parameters and regressors is no longer a problem. In addition, the distributional and spatial autocorrelation assumptions are also now consistent.

The Specification Problem. The hypothetical example violates

[3] Basic EI is the model that was described earlier and is run when the data are inserted into the EI program and no model options are changed. Some options are available to change things such as the maximum number of iterations, step length, or the method of computing area under a distribution. Given convergence, these options should not significantly affect the results. The options that have a direct impact on the value of the parameter estimates are encompassed in extended EI. These options will be discussed extensively later. All results reported for EI in this article were obtained from the EzI program v.1.21 (11/6/96 release).

[4] Extended EI also includes a nonparametric version of the model, as well as other options for setting priors on the covariates, constraints, and computational methods. "The EI model" is the model that is embodied in Chapter 16: "A Concluding Checklist" (King 1997). Points from the checklist will be described throughout the discussion of the extended model.

all three assumptions of the basic model. However, as discussed in Ansolabehere and Rivers (1997) and as suggested in King (Chapter 9), the crucial assumption concerns the correlation between the parameters and the regressors, i.e., aggregation bias. It is possible for the parameters to vary but still not be correlated with the regressors. In these cases, aggregate data models will fare well. Clearly, then, the aggregate data problem can be seen as a specification problem. Achen and Shively (1995) show that the specification problem is not solved simply by using a regression model specification that would be correct for individual-level data. If one can identify subsets of the data where the parameters do not vary, then one can eliminate the problem of aggregation bias. Constancy will exist within the subsets of data. However, identifying the variables that will yield this propitious situation is extremely difficult. The extended EI model allows the addition of covariates for precisely this purpose. In the hypothetical example, partisan leaning was the necessary covariate. When partisanship is controlled, the parameters are constant and all of the assumptions are fulfilled.

In practice, choosing these control variables is the crucial and most difficult part of aggregate data analysis. The aggregate data problem is not solved by determining that additional variables need to be included but, rather, by including the *correct* additional variables. EI provides some diagnostics which purportedly aid a researcher in determining a proper model specification by signaling deviations from the model's assumptions. King claims that "valid inferences require that the diagnostic tests described be used to verify that the model fits the data and that the distributional assumptions apply" (King 1997, 21). I now turn to analyzing how well the diagnostics are able to verify the appropriateness of the assumptions of the model.

The Extended EI Model

Certainly, a researcher needs to know if a proposed model fits the data and if the necessary assumptions apply. Clearly, if the data do not meet the aggregation bias assumption, the model needs to be modified. Indeed, EI is not meant to apply to every possible aggregate data problem (King 1997, 158). In our hypothetical model, the assumptions of basic EI did not fit the data, though modifying the model by including partisan leaning corrected this shortcoming.

The important question to ask is, with real aggregate data where uncertainty abounds, how well are we able to assess whether the specification is correct and whether the assumptions fit the data? If the assumptions do not fit the data, will we be able to determine how to modify the model correctly? Are the EI diagnostics enough to unveil

improper assumptions and to guide us to an appropriate model? In this section, EI along with these diagnostics is tested on two sets of real data.

Data Set 1. The first data set was derived from a survey conducted for the 1984 California general election by Bruce Cain and D. Roderick Kiewiet.[5] In total, the survey has 1,646 respondents and includes an oversampling of ethnic minorities. The data were aggregated into 30 precincts. Since the data are at the level of the individual, the estimates from the aggregate data models can be assessed against the true values.

In this example, the goal is to predict the percentage of college graduates by race based solely on the aggregate data. The accounting identity is

$$\begin{aligned}&(\%\ \text{COLLEGE EDUCATED})\\ &\quad = (\%\ \text{BLACK})\beta^B + (1 - \%\ \text{BLACK})\beta^W.\end{aligned}$$

The known information for this problem is summarized in Figures 3 and 4. Figure 3 is simply a scatterplot of precincts with (% *BLACK*) on the horizontal axis and (% *COLLEGE EDUCATED*) on the vertical axis. Figure 4 is the diagnostic tomography plot for the data with β^B on the horizontal axis and β^W on the vertical axis.[6] For each precinct, there are four quantities of interest. Two of these quantities, the (% *BLACK*) and (% *COLLEGE EDUCATED*), are known while the other two, the percentage of college-educated blacks, β^B, and the percentage of college-educated whites, β^W, are unknown.

Each point in Figure 3 maps to one point in Figure 4, giving us all four quantities of interest. The problem is that the mapping is unknown. Any percentage of blacks and any percentage of whites could be college-educated. However, using the accounting identity and the method of bounds collapses the space of possible (β^B, β^W) values from the whole space to a single line for each precinct (King, Chapter 6). In Figure 4, the true values of (β^B, β^W) are marked by points on the lines. In practice, all we know is that the true value lies *somewhere* along the line. The problem thus can be rephrased as follows. We know that the true (β^B, β^W) value lies along a line. How do we determine where along the line it lies? All of our deterministic information has been used to

[5] Details on the sample can be found in Cain, Kiewiet and Uhlaner (1991).

[6] Achen and Shively (1995, 207–210) originally suggested the idea of graphing the Duncan-Davis bounds and discussed some features of the resulting plots. Achen and Shively observed, "The basic Duncan-Davis limits, aggregated across all the districts' equations, define the outer limits of any possible solution space" (1995, 208). King applied such plots to real data, likewise observing that "the bounds and the lines in this figure give the available deterministic information about the quantities of interest," and called the result a "tomography plot" (1997, 80–82).

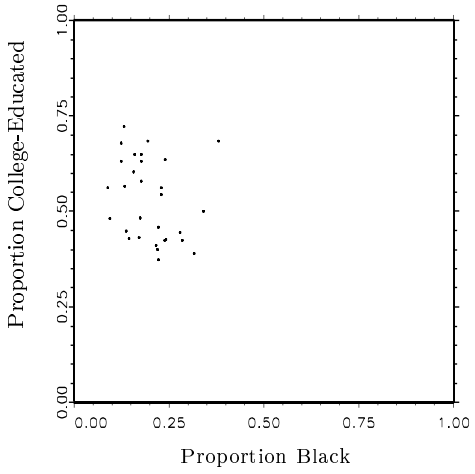

Fig. 3. Data Set 1. Scatterplot of the aggregate data quantities

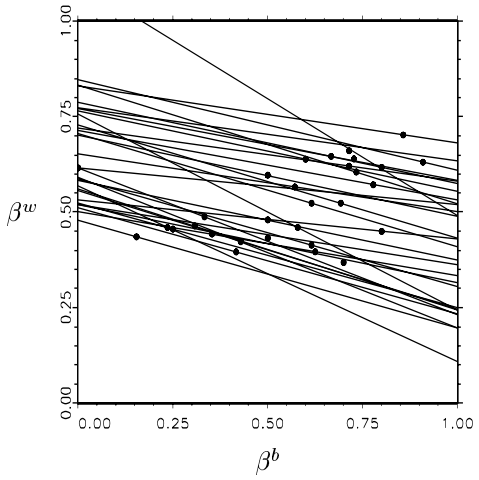

Fig. 4. Data Set 1. Tomography plot with true values plotted

determine the lines in the tomography plot. To make any claims about where the true value lies along the line, we must make some assumptions, which may or may not be true.

With the OLS model, the constancy assumption is made. With respect to the tomography plot, the assumption amounts to the claim that all of the lines should intersect at one point. The extent to which they do not intersect at that point is simply attributed to error. As we can see from Table 4, the OLS estimates are not particularly close to the truth and do not yield good substantive analysis of the problem.[7] The point estimate for whites is closer to the truth than the point estimate for blacks, but it is still more than a standard error away. The problem is that the percentages of college-educated blacks and whites are far from being constant across precincts. In truth, the percentage of college-educated blacks varies from under 1 to 90 percent depending upon precinct. The percentage of college-educated whites does not differ as widely but varies considerably nonetheless, running from 36 to 70 percent. Obviously, the assumption of constancy is likely to be troublesome here.

EI makes different assumptions for determining where the true values lie along the tomography lines. In particular, the assumptions of the basic model will place the point estimate near the greatest density of lines. This process is referred to as "borrowing strength" from other observations to determine the true value for any given observation. In essence, the model does not assume that there should be a single point of intersection but it does assume that all of the lines should substantially intersect in one common area. Implicit here is the unimodality assumption: all of the lines are related to one common mode. Unfortunately, this distributional assumption is also misplaced. An examination of the individual-level data reveals that there are two groups of lines, i.e., that two modes, not one, exist in the data. These two groups of lines can be distinguished as one group which is associated with high-income precincts and another group that is associated with low-income

[7] In these examples, "Truth" is actually an estimate of a true population parameter. The estimate is based on a sampling from a population and thus the "truth" has a sampling error component to it. However, in these examples, the respondents in the survey will be considered the population universe. Hence, no standard errors are reported for the "Truth." The numbers are simply an accounting of the data. This is, in fact, the setup of both EI and Goodman. Neither EI nor Goodman incorporate the sampling error into the model. Both models assume that the marginals are known. This assumption can be very influential in samples where the sample size is small. To distinguish between sampling error and standard errors from the model, the phrase "Model standard errors" is used. All values in the table are standard errors from the model and do not incorporate sampling error.

TABLE 4. Predicting Education Level by Race

	Black	White
Truth	.5343	.5126
OLS	.2322	.6042
	(.2223)	(.0584)
Basic EI	.3404	.5747
	(.3993)	(.0845)
Extended EI	.4904	.5399
nonparametric version	(.0471)	(.0117)
Extended EI	.5060	.5360
covariate: Income	(.0551)	(.0137)
Extended EI	.1660	.6207
covariate: Age	(.3220)	(.0802)

Model standard errors in parentheses.

precincts. Since the two sets of lines are not related to the same mode, we would not want to "borrow strength" from one group of lines to determine the mode of the unrelated other group of lines.

In addition, as in the hypothetical example earlier, viewing the individual-level data reveals that aggregation bias also exists. Indeed, while correlation of the parameters and regressors does not follow from varying parameters, these two situations will commonly occur in tandem in aggregate data. King acknowledges that " ... it pays to remember that most real applications that deviate from the basic ecological inference model do not violate one assumption while neatly meeting the requirements of the others" (King 1997, 159). With regard to this particular data set, one should not expect the basic EI model, with its erroneous assumptions, to provide particularly good estimates. And, indeed, as we can see from Table 4, basic EI reports similar point estimates and comparable standard errors to OLS. After accounting for the standard errors, the point estimates are statistically indistinguishable.[8]

Since the basic EI model makes the wrong assumptions about the data, we should expect some indication of this through the diagnostics. The tomography plot in Figure 5 is the suggested diagnostic for determining modality. Here, we should find evidence of multiple modes in the data. A mode is indicated by a mass of lines, preferably intersecting lines. So two distinct groups of lines would indicate two modes. However, in our tomography plot, there is no evidence of multiple modes.[9]

[8] In general, one would not expect OLS and EI estimates to be much different. The reasoning is that when the assumptions of OLS are violated, the assumptions of EI are also violated. The point at which the models diverge is when the OLS estimates are beyond the bounds. In these instances, EI may provide better estimates.

[9] One should be cautioned that there is considerable uncertainty encompassed in

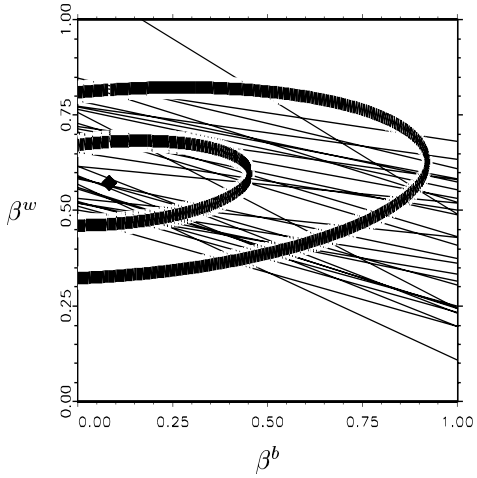

Fig. 5. Data Set 1. Tomography Plot

If multiple modes did exist, we would want either to add covariates to isolate the multiple modes or to use the nonparametric version of the program to bypass the assumption of truncated bivariate normality. Properly accounting for the modes will simultaneously solve the aggregation bias problem. On the other hand, according to the logic conveyed in "the checklist" (King 1997, Chapter 16), we note that the truncation of the contour lines in the tomography plot is fairly heavy, and that this is supposed to provide confidence that the basic model should be fine (King 1997, 284). In addition, the results also do not seem substantively unreasonable, and this also supposedly provides credence for the basic model. Based upon the diagnostics and reasoning suggested in "the checklist," then, the basic model should be appropriate. Only our knowledge of the truth tells us otherwise, that the basic EI model has led us astray.

Suppose, however that the researcher did believe that the results were not correct, for one reason or another. Perhaps the researcher believes that aggregation bias exists. Figure 6 is suggested in the Checklist (item 10) as an indicator of aggregation bias (King 1997, 283). This plot

---

a search for multiple modes whether this search is through tomography plots or the nonparametric density plot. If one really wanted to see multiple modes in Figure 5, one could probably convince oneself that they exist. To boot, the nonparametric plot might even provide some supporting evidence in this vein. In this case, multiple modes do exist so one then needs to resolve the conflicting nature of the tomography plot and the nonparametric plot. In other plots (e.g., see King, Figure 9.1a), many modes will seem to exist in the tomography plot when, in fact, only one mode exists.

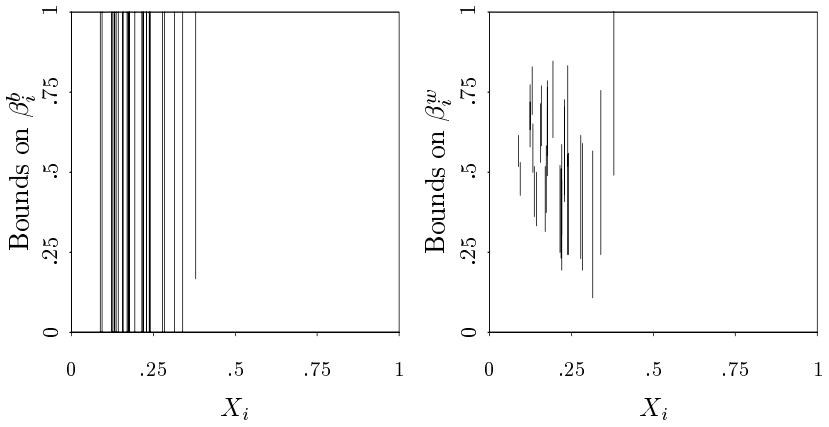

Fig. 6. Data Set 1. Aggregation Bias Diagnostic

gives us some indication of the possible correlation between the Xs and the βs. The true β values are unknown and lie on some unknown position along each line. Judging from these plots, aggregation bias may exist or it may not exist. Especially for β^b, neither conclusion is warranted though both are possible and both hypotheses can be supported by substantive beliefs. The pattern for β^w might be increasing, implying a correlation, or it may be random, implying no correlation. This diagnostic clearly has very limited utility in this application.

If one believes that aggregation bias does exist, one might try including certain covariates to alleviate this problem. Certainly one could make a credible argument for including income as a covariate. Income clearly affects education, and it is well known that the two variables have a strong relationship. Alternatively, one could make an equally credible argument for including age as a covariate. Clear evidence exists that the American population has become significantly more educated over time. Either of these two scenarios is reasonable based on qualitative information and substantive beliefs.

The results of the extended EI models with these covariates are reported in Table 4. As we can see, extended EI with income as a covariate does fairly well. Note, though, that extended EI with age as a covariate produces significantly different results. An obvious problem with adding covariates is that King provides no method of choosing covariates (outside of utilizing qualitative information and substantive beliefs). However, selecting the proper covariates, or determining the proper specification, is the heart of the problem. Using the type of qual-

itative information and substantive beliefs suggested in King's Chapter 16 is plainly inadequate. These criteria are subjective and can differ wildly between researchers. As our example indicates, these beliefs can substantially affect our results yielding different point estimates and different estimates of uncertainty. Disturbingly, the resulting substantive claims from different models are inconsistent. Given that this is the case, how would a researcher choose one model's result over another? How much faith should be placed on qualitative beliefs? Believing that the data should be separated by income does not mean that income provides a true demarcation of the data. Formal and objective tests are necessary. Visual examinations of the tomography plots are simply not good enough.

Lastly, one might consider the nonparametric version of the model since it does not depend on such a rigid distributional assumption. Unfortunately, it seems evident that in this case the standard errors from the nonparametric model are erroneous. The true proportion of college-educated whites differs from the nonparametric point estimate by more than two standard errors. The nonparametric estimates are less precise than the reported standard errors would suggest they are. How much less is not clear. Worse, the diagnostic plots have not given us a clear reason to pursue this model over the other models or to pursue the other models over the nonparametric model.

The important question to ask here is, which model would we have chosen if we had not had the individual-level results on hand? In which model are the assumptions correct? The diagnostic plot in Figure 5 did not indicate a bad model fit. Hence, the researcher might well feel justified in reporting the results of the basic model along with a comforting note that the diagnostics verified the adequacy of the model. Basic EI in reality did no better than OLS with its assumption of parameter constancy. In addition, extended EI utilizing covariates did not converge on one clear and consistent answer.

Lastly, we note that the estimates from the basic EI model are not "incorrect." Like OLS, the estimate for β^W is fairly accurate with a fairly small standard error. Also similarly, the point estimate for β^B, while not near the truth, correctly indicates a large degree of uncertainty. These assessments are extremely useful. However, neither OLS nor the basic EI model provides a good substantive understanding of the data. Interpretation of both models would lead one to believe either that we can make no comparisons between educational levels or that the level of education among blacks is not likely to be as great as it is among whites. In fact, the rates are almost identical in this data set. In addition, while adding income as a covariate corrects the model, there is no clear

TABLE 5. Predicting Vote for Thomas Hsieh by Race

	Chinese	Non-Chinese
Truth	.8849	.6300
OLS	.9069	.4890
	(.0338)	(.0674)
Basic EI	.8573	.6061
	(.0628)	(.1483)
Extended EI	.8539	.6143
nonparametric version	(.0114)	(.0270)

Model standard errors in parentheses.

indication that this model should be believed over the others or that this approach is more credible than the model that erroneously, as it turned out, included age as a covariate. Our problem now is that we have a set of "believable" models which yield an array of "solutions" but no clear way of distinguishing good models from bad models.

Data Set 2. The second example examines data which were compiled by Larry Tramutola and Associates and include a total of 1497 respondents in 37 precincts. The target population was Asian Americans in Northern California. The goal is to find a model which accurately estimates the support for Thomas Hsieh in his bid for San Francisco City Council. A researcher interested in his support among Chinese voters who had only the aggregated precinct data on votes and ethnicity might estimate an OLS model as

$$(\%\ \text{HSIEH VOTE}) = (1 - \%\ \text{CHINESE})\beta^N + (\%\ \text{CHINESE})\beta^C + e.$$

Again, OLS and EI are tested, and the EI diagnostic tools are examined.

OLS results are reported first in Table 5. As we can see, these results are good. Next, the EI estimates are reported. These results are also quite good. Due to the large standard errors, in fact, the basic EI estimates are again statistically indistinguishable from the OLS estimates. The assessment of the results as good, of course, depends on our full knowledge of the truth—a luxury never accorded in practice. Hence, to assess our models more realistically, momentary ignorance of the truth is feigned.

Observe the model diagnostics. The tomography plot is displayed in Figure 7. A considerable amount of interpretative leeway is available in assessing the plot. However, this figure does not display any striking evidence suggesting multiple modes or aggregation bias. One could argue for the basic EI model on the basis of the heavy truncation of the

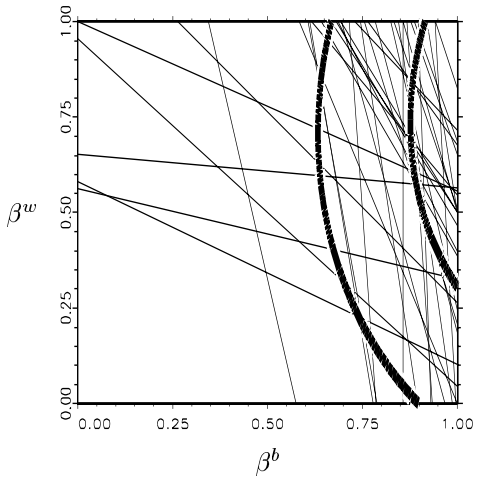

Fig. 7. Data Set 2. Tomography Plot

contours in the tomography plot, lack of evidence for multiple modes, or a reasonable estimate from Goodman's regression. On the other hand, extended EI might be necessary based on alternative qualitative beliefs, a substantive sense that other variables may matter, or an uneasiness with the distributional assumption.[10] However, in this example, it is of no consequence to the point estimates whether a single mode model or a multiple mode model is chosen. The parametric model and the nonparametric model give almost identical point estimates. The difference is that the nonparametric model expresses substantially less uncertainty. The ostensible precision is, in fact, misleading—the nonparametric point estimate for Chinese voters is in reality more than two standard errors from the true parameter value.

In this example, as in the two previous examples, it is difficult to pinpoint the most reasonable option. While the point estimates are similar, the measures of uncertainty vary quite a bit. Is it most reasonable or safest to assume the least amount of certainty? As always, the model that is the most reasonable is the one in which the assumptions made are most true. However, we are unable to make a good assessment here of which model encompasses the most reasonable set of assumptions. Since the diagnostics have limited usefulness, we are left with only our qualitative beliefs to guide us. Again, we are presented with a set of equally believable models and no method for choosing between them.

[10] King provides a list of these "tests" in the checklist (1997, Chapter 16). In addition to simple qualitative beliefs, they include a scattercross plot, an examination of the bounds and the contours, Goodman's regression line, and the tomography plot.

Assessing EI

What do these examples suggest? EI represents a genuine advancement to ecological inference in that it incorporates two elements that have never previously existed together in aggregate data models. The combination of the method of bounds and allowance for varying parameters brings a new degree of efficiency to aggregate data analysis. The point at which caution is essential, however, is when the assumptions of the model are inconsistent with the data. To a limited extent, one can gauge the suitability of EI's assumptions for the data by the model diagnostics. Hence, the model diagnostics should be used *every* time EI is employed. However, the diagnostics are problematic in that they do not always signal deviations from the model even when they do exist. Alternatively, the diagnostics sometimes point toward a poor model fit when the estimates are actually quite reasonable. In addition, the diagnostics are based on visual assessments and substantive beliefs—two elements which can be completely random but equally believable across researchers.

EI is appropriate if and only if the specification is correct, i.e., if and only if there is no correlation between the parameters and the regressors. The problem is that one has no idea whether the specification is correct or not, and the diagnostics have limited utility in this regard. EI does not bring one much closer to knowing the underlying structure of aggregate data. EI merely provides a program through which one can construct many different model specifications. Herein lies a fundamental and serious weakness of EI. Intuition can bring us only so far. Without a formal method for determining how to extend the model, a researcher is left with a wide variety of "reasonable models" and no way of assessing whether *any* of these models is appropriate.

Clearly, aggregate data models should have formal diagnostics to help determine a proper specification. Covariates should be chosen on the basis of the properties of well-known statistics rather than intuition or qualitative beliefs. In particular, since the addition of different covariates may significantly affect the resulting point estimates and measures of uncertainty, it would be useful to have a measure which assesses the likelihood that a covariate distinguishes between distinct subsets in the data and thus alleviates the problem of aggregation bias. Such a statistic is provided in other aggregate data models (Cho 1997). These statistics are useful when employed in a switching regimes context or as an added component in the context of EI.

In summary, a caveat is implored. Caution should never be thrown to the wind in ecological inference. Never should a model be run without a full understanding of the implications of its assumptions. No model

should be treated as a black box solution to the aggregate data problem. As with any model, EI is built upon assumptions, and these can be far off or right on target. The estimates therefore may also be far off or right on the true parameters. Substantive discussions of the results of EI should thus always include a discussion of the assumptions, how reasonable they are for the problem at hand, and how these assumptions drive the results. Excitement about the advances to ecological inference provided by EI should not be allowed to lead to insufficient attention to the strong and potentially inappropriate assumptions at the heart of the model. The model is useful if and only if the assumptions fit.

REFERENCES

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

Ansolabehere, Stephen, and Douglas Rivers. 1997. "Bias in Ecological Regression Estimates." Working paper.

Berger, James O. 1985. *Statistical Decision Theory and Bayesian Analysis.* 2nd ed. New York: Springer-Verlag.

Box, George E. P. 1953. "Non-normality and Tests on Variances." *Biometrika* 40:318–335.

Cain, Bruce, D. Roderick Kiewiet, and Carole J. Uhlaner. 1991. "The Acquisition of Partisanship by Latinos and Asian Americans." *American Journal of Political Science* 35:390–422.

Cho, Wendy K. Tam. 1997. "Structural Shifts and Deterministic Regime Switching in Aggregate Data Analysis." Working paper.

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

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

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

Huber, Peter J. 1981. *Robust Statistics.* New York: John Wiley & Sons, Inc.

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

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

Scheffé, Henry. 1959. *The Analysis of Variance.* New York: Wiley.

Shively, W. Phillips. 1974. "Utilizing External Evidence in Cross-Level Inference." *Political Methodology* 1:61–74.

Swamy, P. A. V. B. 1971. *Statistical Inference in Random Coefficient Regression Models.* Berlin: Springer-Verlag.