Geography, Statistics, and Ecological Inference

Gary King
Department of Government, Harvard University

I am grateful for such thoughtful reviews from these three distinguished geographers. Fotheringham provides an excellent summary of the approach offered, including how it combines the two methods that have dominated applications (and methodological analysis) for nearly half a century—the method of bounds (Duncan and Davis 1953) and Goodman's (1953) least squares regression. Since Goodman's regression is the only method of ecological inference widely used in Geography (O'Loughlin), adding information that is known to be true from the method of bounds (for each observation) would seem to have the chance to improve a lot of research in this field. The other addition that EI provides is estimates at the lowest level of geography available, making it possible to map results, instead of giving only single summary numbers for the entire geographic region. Whether one considers the combined method offered "the" solution (as some reviewers and commentators have portrayed it), "a" solution (as I tried to describe it), or, perhaps better and more simply, as an improved method of ecological inference, is not important. The point is that more data are better, and this method incorporates more. I am gratified that all three reviewers seem to support these basic points. In this response, I clarify a few points, correct some misunderstandings, and present additional evidence. I conclude with some possible directions for future research.

Ecological Inference as Statistical Inference

John O'Loughlin argues that "the specific problems of geographic data analysis require a different mode of thinking than is usually found

in inferential statistics" (p. 595). Although I recognize the place-space divide he is referencing here, inferential statistics should not be seen as a threat to anyone interested in learning about the world. It is fully general and capable of evaluating any logically consistent statement with observable implications. No other approach dominates it, even when only qualitative information is available (King et al. 1994). In this section, I address this point in the context of ways the reviewers have characterized the ecological inference problem.

What Is Inference?

Anselin correctly points out that the quantities of interest in ecological inference are not observable. Although they were originally observable, these quantities have been lost in the process of aggregation. Unfortunately, he greatly confuses matters when he describes ecological inference as "creat[ing] data where no data exist" (p. 587). In fact, all problems of statistical inference share exactly this feature.

To be specific, inference is the process of using facts we know to learn about facts we do not know.[1] In ecological inference, the facts we do not know are the contents of the cells in a set of contingency tables. In forecasting, the facts we do not know are future values of our outcome variable. In descriptive inference, the facts we do not know are population parameters (which we would seek to learn using observations taken from some type of sample). In causal inference, the fact we do not know is the average difference between the dependent variable for a unit (respondent, country, etc.) when a "treatment" is applied and the same unit when a "control" is applied, the problem being that only one value of the dependent variable is observed for any unit (this is known as the Fundamental Problem of Causal Inference; see Holland 1986; King et al. 1994).

In another context, Anselin (1988) artfully reviews a variety of methods of making inferences about spatial autocorrelation parameters. Since no one has ever observed a spatial autocorrelation parameter in the real world, we could equally describe his methods as "creating data where none exist," being "essentially unverifiable" (p. 587), "observationally equivalent" (p. 588) to other contradictory models, "not a solution to the [spatial autocorrelation] problem," or even as estimating quantities that may not exist. Language can be fun, but none of this should take away from the very real contribution these models make in enabling scholars to draw inferences about unknown spatial autocorrelation parameters from known observations of geographically coded variables. In sum, models of ecological inference are not unique in helping us learn about facts we do not know. Whether this is "creating data where none exist" is a small semantic point. Whatever language one chooses, it applies to all types of inference.

Evaluating Model-Based Inferences

Ecological inference and most other statistical analyses in the social sciences are model-based, meaning that inferences depend on data as well as statistical assumptions. This means that, Anselin's claims notwithstanding, there are indeed several "yardstick[s] to compare competing methods" (p. 587) of ecological inference. They are the same yardsticks used for virtually all statistical methods. For starters, we can determine whether estimates have attractive statistical properties when the assumptions are correct. Anselin seems to apply this approach when he concludes that "there is no overarching statistical framework to encompass these methods (such as asymptotics, or normal distribution finite samples)" (p. 588). He also recognizes, in a slightly different context, that the method "is firmly grounded in modern statistical concepts familiar in the literatures dealing with Bayesian hierarchical modeling, . . ." (p. 590). In fact, the latter statement, which is correct, can be used to correct the first, which is not. That is, Bayesian models like EI have the full range of desirable statistical properties. For example, aggregate estimates conditional on the assumptions are admissible, statistically consistent, asymptotically normal, and asymptotically efficient (the same as, say, logistic regression or probit). The extensive Monte Carlo evidence presented in the book demonstrates that estimates are approximately unbiased and efficient in samples as small as twenty-five (something not true of logistic regression, for example). The model can also be justified by appeal to likelihood theory, since priors are not necessary.

A second way to evaluate model-based inference is to evaluate the properties of the estimators when the assumptions are wrong. As

O'Loughlin and Fotheringham point out, my book contains a forty-page section about what can go wrong with these assumptions. The conclusion from this exploration by me and others is that the method is more robust to incorrect assumptions than Goodman's regression, but the degree of robustness is specific to the data chosen. Sometimes including the information from the bounds provides enormous amounts of additional information; other times the bounds do not add much. Fortunately, we can often tell which situation we are in, and there is little chance of doing worse by using additional information.

Third, observable implications of the model can be checked to assess the fit. Although information is lost by the process of aggregation, some observable implications do indeed exist and can be used to rule out some models. It would be nice if there were more (just as it would be nice if there were no information lost to aggregation), but at least there are some, and they should be used. Indeed, the book describes, and the software implements, many diagnostics that take advantage of these observable implications. I think all the reviewers recognize the value of these kinds of diagnostics, though Anselin worries that they require "nonrigorous visual interpretation" (p. 591). The book describes and the software includes formal hypothesis tests for aggregation bias (based on the coefficient on X in the mean function), but I do prefer graphical diagnostics. This is in part a matter of taste: Anselin's view is that formal statistical tests are preferable since the decisions drawn are more clear-cut, and because of the problems uncovered in the literature on cartography and perception. Others prefer graphical diagnostics, however, since they convey more information, consider a larger range of observable implications, and have the ability to reveal features of the substantive problem and data we did not realize to look for in the first place. Of course, graphical tests are less precise and leave more open to interpretation, and so graphics and statistical tests can both be useful tools. As a general matter, graphical tests are better when we know less; formal statistical tests are better when we are confident of all aspects of the model other than the one being tested (and are invalid otherwise). For ecological inference, I would prefer to assume we know less, but I would welcome any development of other formal theories.

Ecological inferences involve nothing mysterious or even unusual from a theoretical statistical perspective. They share the same characteristics as all other model-based inferences. What makes ecological inference an especially hazardous process is that we do not always have sufficient external information to be comfortable with the assumptions of the model. The fix, for ecological inference and all other such inferences, is to gather this information. Fortunately, geographers are in an especially good position to do so.

Consequences of Ignoring Spatial Autocorrelation

In my book (x 9.3.1), I found that spatial autocorrelation had minimal effects on estimates and standard errors from the model. This result was found to be "particularly interesting" by Fotheringham (p. 585) but the method by which I arrived at this conclusion was questioned by the other two reviewers because I evaluated the model under a relatively simple version of spatial autocorrelation. As a result, O'Loughlin writes that "King's spatial dependency simulations are unconvincing." Anselin agrees and explains that the right Monte Carlo experiments should include "the simultaneity induced by multidirectionality and two-dimensionality (Anselin 1988). It remains to be seen how real spatial autocorrelation would affect the results of EI" (p. 590).

These are reasonable criticisms, and I am grateful for the opportunity they present to expand on the properties of EI. For this response, I ran another Monte Carlo experiment according to the specifications in Anselin (1988). In order to ensure a realistic form of multidirectionality and two-dimensionality, I used a spatial contiguity matrix coded from all nations in the world (with at least one neighbor). So as to avoid confounding, the data were generated so that the other (distributional and correlational) assumptions of the model were satisfied. I used the same number of simulated datasets (250) as in my book.

The results, reported in Table 1, confirm the conclusions from the simpler analysis presented in my book: Spatial autocorrelation has only a minimal effect on model estimates and standard errors. For example, the average absolute error of the aggregate (global) quantity of interest is approximately zero for both the independently

Table 1. Consequences of Spatial Autocorrelation: Monte Carlo Evidence

	Aggregate-Level			Precinct-level	
Model	Error	(S.D.)	Avg. S.E.	Error	(S.D.)
Independent	.007	(.011)	.014	.01	(.05)
Spatial	.008	(.011)	.014	.01	(.04)

Each row summarizes 250 simulations drawn from a model with areal units that are either independent or spatially autocorrelated, as described in the text. The aggregate-level columns report the average absolute difference between the truth and estimates from the model. The precinct-level columns give the deviation from 80 percent in coverage of the 80 percent confidence intervals (in both cases with standard deviations across simulations in parentheses).

generated and spatially autocorrelated data. The true standard deviation across the 250 simulations (given in parentheses) is fairly close to the average of the estimated standard errors from each simulation, indicating that the aggregate uncertainty estimates are reasonably accurate in this case. In addition, the average error in covering the true values for the 80 percent-confidence intervals is very small for the simulations both with and without spatial autocorrelation.

Since information is lost in the process of aggregation, using all information in the data and adding additional qualitative, geographical, and other external information is the only way to be reasonably confident when making ecological inferences. I welcome the suggestions of all three authors about additional geographical information we might include in the analysis to further improve our inferences. I would welcome attempts to extend the model offered to incorporate this information, along with analyses like the one in Table 1 that verify whether or not they make a difference. The additional evidence offered here supports my conclusion in the book that spatial autocorrelation has minimal effects on other aspects of this model.

Future Directions

The time has never been better for developing improved methods of ecological inference. More scholars from more disciplines are working on improving and applying methods of ecological inference than ever before. As a result, more methodological progress has been made in the last few years than at any time since Goodman and Duncan-Davis were writing. The most productive future paths will almost surely be those that bring in the most quantitative and qualitative knowledge external to the data at hand, and it is hard to think of a discipline with more of this contextual knowledge than geography. I encourage geographers to bring their skills to bear on this important question for their own research and for those in numerous other disciplines and nondisciplinary areas.

For example, John O'Loughlin approaches the problem from one angle by writing that "Only careful survey research that incorporates specific contextual questions will resolve the political science-political geography difference of opinion on the nature and significance of context" (p. 600). I agree about the benefits of including aggregate and contextual information in public-opinion surveys (and have indeed argued for this position in other contexts; see King 1996), but this is only half the story. That is, survey research is a very powerful tool, but even the best surveys will not resolve the problem alone. For example, I doubt anyone would want to mount a survey in present-day Germany to ask survivors whether they voted for Hitler in the 1930s and 1940s! Indeed, even in the circumstances where survey responses are most likely to be sincere, no survey includes enough respondents to provide good estimates for local areas.

The other half of the story is to use survey data to improve methods of ecological inference. There has been some other work on this (e.g., Little and Wu 1991), but important opportunities remain. In particular, what surveys are good at, in practice, is providing a good snapshot of averages over an entire country. Fortunately, this type of information is precisely what we need in order to relax some key methodological assumptions in making ecological inferences. For example, even a relatively small survey can greatly improve the estimation of the five key parameters in the basic version of my model, or the degree and direction of aggregation bias, and once we include this information, estimating all the precinct-level parameters is far less dependent on assumptions.

A second fertile area for future work is modeling the special types of measurement error in ecological data. Anselin raises one possibility when he expresses concern about "sampling error associated with T_i" and suggests that we model this with binomial random variables. As it turns out, this has been tried in the literature, in Brown and Payne (1986) and King et al. (1999), from which it is clear that including binomial variation is substantively meaningless in most realistic political geography applications. For example, consider the variance of Ti for an average-sized electoral precinct (the smallest unit of political geography in the U.S.). Such a precinct (with, say, 1000 people), evaluated at 0.5 probability (where the variance is largest), has a variance of only 0.00025. Of course, this should not be surprising: sampling error is mostly a problem in sample surveys. Although aggregate data has far less sampling error, many other types of measurement error need to be modeled. For example, it would be very useful to develop models for errors induced when matching two sets of data on overlapping geographies, or due to mismatched geocoding. As is, these are dealt with only by cartographers and data providers and then ignored during statistical analyses. In fact, these are critical parts of the data generation process but have not been modeled statistically in the context of ecological inference.

The reviewers' suggestions of using geographically weighted regression, spatial econometrics, and spatial expansion are also promising general approaches, and useful for many specific problems, but of course they cannot be used without modification to make inferences about individuals from aggregate data. In the spirit of these suggestions, however, I experimented with several extensions of the nonparametric method discussed in the present work (x 9.3.2) with kernels defined on the basis of geographic rather than tomographic proximity. I have no doubt that this will help in some cases, but I could not find a real example with political data where it noticeably improved the inferences. This would seem to be another productive avenue to follow.

Other issues worth further exploration include methods of model selection, exible specifications, methods for continuous individual-level variables, and explicit models of spatial autocorrelation. Ori Rosen, Martin Tanner, and I have been working on extensions to multiple category variables, Philip Cross and Charles Manski appear to be making progress on extensions to continuous variables, and a dozen others have written articles pushing forward other aspects of the problem. As Fotheringham reemphasizes, further work developing aggregation-invariant statistics will provide additional help in avoiding the Modifiable Areal Unit Problem in new areas. To further encourage this research, I make the following offer: Develop a better method of ecological inference—a test, diagnostic, or other related tool—demonstrate its value, and provide some code, and I'll include it in the *EI* and *EzI*, the widely used software packages that implement these methods (available at http://GKing.Harvard.edu).

Acknowledgment

My thanks go to Luc Anselin, Stewart Fotheringham, and John O'Loughlin for helpful comments and conversations.

Notes

1. In Bayesian analysis, for example, the only two types of quantities that exist are those that are known or unknown. Known quantities are fixed numbers; unknown quantities are modeled via probable densities.
2. Precinct-level parameter estimates in EI have the same properties as the aggregate estimates, except of course that their posteriors do not collapse asymptotically (for the same reason that the nonzero variance of −, ÿ 2, prevents the distribution of y = x _ + _ from collapsing asymptotically in linear regression). A related confusion is Anselin's claim that since EI has "2N parameters but only N observations, estimation is clearly impossible" (p. 589). This is an intuitive-sounding claim, but it does not follow in ecological inference or in other statistical models. For example, in least squares regression, with k explanatory variables (including the constant), there are k + 1 parameters (i.e., the _s and ÿ 2). But from these we can estimate any number of conditional expectations (using fitted values from the regression). Similarly, the basic version of EI has only five parameters, from which we can easily estimate two (and indeed many more, if desired) observation-level parameters. Anselin's misstatement is explained by the fact that one can estimate at most, N independent parameters from N observations, whereas precinct-level quantities of interest in ecological inference (and conditional expectations in least squares regression) are not independent: in fact, conditional on the data and one precinct-level

parameter, the other parameter may be computed deterministically (see King 1997: 80, Equation 5.2).

References

Anselin, Luc. 1988. *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers.

Brown, Philip J., and Payne, Clive D. 1986. Aggregate Data, Ecological Regression, and Voting Transitions. *Journal of the American Statistical Association* 81:452–60.

Duncan, O.D., and Davis, B. 1953. An Alternative to Ecological Correlation. *American Sociological Review* 18:665–66.

Goodman, Leo. 1953. Ecological Regressions and the Behavior of Individuals. *American Sociological Review* 18:663–66.

Holland, Paul. 1986. Statistics and Causal Inference. *Journal of the American Statistical Association* 81:945–60.

King, Gary. 1996. Elections and the National Election Studies. Paper prepared for the National Election Studies, Congressional Elections Conference, copy available at http://gking.harvard.edu/preprints.shtml≠nes.

———; Keohane, Robert O.; and Verba, Sidney. 1994. Designing Social Inquiry: Scientific Inference in Qualitative Research. Princeton, NJ: Princeton University Press.

———; Rosen, Ori; and Tanner, Martin A. 1999. Binomial-Beta Hierarchical Models for Ecological Inference. *Sociological Methods and Research*, in press.

Little, Roderick J.A., and Wu, Mei-Miau. 1991. Models for Contingency Tables with Known Margins when Target and Sampled Populations Differ. *Journal of the American Statistical Association* 86:87–95.