Isolating Spatial Autocorrelation, Aggregation Bias, and Distributional Violations in Ecological Inference: Comment on Anselin and Cho

Gary King
Center for Basic Research in the Social Sciences,
34 Kirkland Street, Harvard University,
Cambridge, MA 02138
e-mail: king@harvard.edu

Few better ways of checking and improving statistical methods exist than having other researchers go over your results, and so I especially appreciate the efforts in Anselin and Cho (2002), hereinafter AC. In this note, I make two main points.

First, AC's numerical findings from its empirical example and simulations contradict no prior research. The article's one empirical example violates EI's spatial independence *and* no aggregation bias assumptions, according to AC, and thus offers no evidence of the independent effects of either. In AC's Table 1, Goodman's regression gives one answer 11.2 times smaller than the truth and logically impossible (the percentage of males having strokes is 6.5% fewer than there are males) and another 8.5 times larger than the truth. Yet, EI gives answers that are 1.41 and 0.74 times the truth. This pattern is common and occurs for a reason: Although aggregation bias can cause Goodman's regression to be biased to any degree, EI's potential bias, although not guaranteed to be zero, is strictly limited, and hence more robust.[1]

Similarly, in simulations with autocorrelation levels set considerably higher than any published ecological inference application, AC's numerical results still confirm that spatial autocorrelation has modest effects. Results in the article's Tables 2 and 3 are similar to and often smaller than those in King (1997, Ch. 9; 2000). As with heteroskedasticity or autocorrelation in linear regression, AC find that EI is unbiased in the presence of spatial autocorrelation, and it has a proportionately larger variance than data without autocorrelation. Because finding autocorrelation is another way of saying the data contain less information, this is precisely as it should be. The only issue is the extent to which EI's uncertainty estimates miss this loss of information, but AC does not address this issue.

My second main point is that AC ignores the role of the bounds in EI and is mistaken when it claims that "setting aside consideration of the role of the bounds does not have a material consequence on [*sic*] our discussion." The advantage of EI comes precisely from combining the only two approaches that had been used in practice prior to EI—Goodman's regression

[1] For a more precise definition of robustness, Goodman's regression has a "breakdown point" (the smallest proportion of observations that would have to be changed to move an estimate arbitrarily far from the truth) of $1/n$, whereas *EI* has a breakdown point of at least 1. See Donoho and Huber (1983) and King (1997, p. 180).

and Davis and Duncan's bounds. AC's omission of the bounds—the most important and only certain source of information—has several serious consequences.

For one, all models of spatial autocorrelation introduced in AC's Section 5 are logically impossible. The error terms cannot be iid, and the functional forms cannot be correct, given the bounds. For another, in trying to use inconsistent models to create consistent simulations, AC offers two modifications. The article's "censored" model heaps observations on T that are outside the bounds at 0 or 1, hence violating EI's distributional assumption. Its "truncated" model discards observations that are not possible, which induces nonrandom selection effects. Because neither approach nor AC's empirical example isolates the effects of spatial autocorrelation, the article contains no information about the assumption's independent effects. (I discuss the way I would draw and have drawn simulations in the Appendix.)

Ignoring the bounds also means the article neglects to mention that its simulations generate highly uninformative bounds, and so should be interpreted as close to a worst case, rather than an empirically representative, scenario. With sufficiently narrow bounds, EI does well no matter what model violations occur. Although Anselin and Cho report not being able to provide me access to their data, it seems clear that the lower (but probably not upper) bounds in their empirical example are highly influential. Ignoring this causes the article to miss the fact that mean posterior estimates, which were used, can be horrible summaries of highly skewed posteriors, as is typical in such applications.[2]

My final main point is that the article misses the point of the EI assumptions and their interrelationships. For one, AC writes that "the spatial aurocorrelation of the dependent variable is not as interesting for our quest here, since it does not necessarily imply spatial autocorrelation of the rates among males (β_m) and females (β_f)." AC's quest, which involves autocorrelation tests of these usually unknown parameters, gets the assumption backward. The assumption is that the dependent variable is independent across observations, conditional on X and any covariates (it is necessary for taking the product over the observations of the dependent variable's density in the likelihood). As a result, AC's tests have no necessary connection to whether the assumption is violated.

The article also gets the definition of the aggregation bias assumption wrong. The article states, "as long as $[X]$ is assumed to be exogenous, there cannot be any aggregation bias." In contrast, the assumption requires that X be unrelated to β_i^b and β_i^w in the "sample." If the assumed relationship is more restrictive than the actual one, regardless of whether X is exogenous, aggregation bias can result.

One consequence is a theme of AC's: that aggregation bias, which is known to have large effects in data that have wide bounds, and spatial autocorrelation have a "clear connection." For example, AC writes, "spatial autocorrelation is symptomatic of aggregation bias." Yet, no mathematical relationship exists: either, both, or neither can occur in a data set. Similarly, no evidence has been offered of an empirical relationship. If a violation of one assumption is detected, we have no more information about whether the other is violated. To understand the isolated effects of one assumption, a controlled experiment, summarizing the consequences of violating only one assumption, is necessary. Such an analysis does not appear in their empirical example or simulations.

[2] In addition, of the many possible EI models, the model selected by AC to present was especially badly misspecified for its data. As the EI manual explains, to avoid bias when analyzing rare events data, several of EI's default options must be changed, but these were ignored in AC. For example, it appears that AC's numerical tolerance parameter was set so that EI rounded many observations to zero preestimation and simulations to zero postsimulation.

Researchers interested in making ecological inferences while taking spatial relationships seriously have only one logically consistent set of models: the EI extended model with variables that tap into spatial features of the data. EI includes formal tests for whether these variables belong in the model (t tests on coefficients and first differences for changes in these variables), graphical diagnostics, and Bayesian model averaging (Imai and King 2002). I would welcome the development of improved methods that tap other spatial features of the data (which I would be happy to include in EI software). The opportunities include improving the performance of EI's uncertainty estimates and, more importantly, discovering patterns not otherwise detectable—both important goals even when spatial autocorrelation does not have a major effect on models that assume its absence.

Appendix: How to Draw Simulations in Ecological Inference

The correct procedure for simulation in ecological inference is to follow some version of these steps, in order: (1) choose values for the parameters and X_i (and any Z_i); (2) draw β_i^b, β_i^w from an explicit, logically consistent model; and (3) compute $T_i = X_i \beta_i^b + (1 - X_i)\beta_i^w, \forall i$, deterministically without selecting or modifying T_i afterward. This automatically satisfies the bounds, almost no matter how steps (1) and (2) are modified for studying particular aspects of the problem.

To follow one of the EI models, draw in step (2) from independent truncated bivariate normal densities with a fixed variance matrix, and with means either fixed or varying as a function of Z_i.

To isolate the effects of spatial autocorrelation, (1) and (2) must respect as much as possible the no aggregation bias [$\mathrm{Corr}\,(\beta^b, X \mid Z) = \mathrm{Corr}\,(\beta^w, X \mid Z) = 0$] and distributional assumptions. In King (1997), I did this by changing Step (2) to use a version of time series dependence in generating β_i^b, β_i^w, which is, of course, a special case of spatial autocorrelation. In King (2000), I set X_i in (1) to values evenly spaced in [0,1] and, to examine the situation of narrow bounds, to [0,0.1], among many other options. For step (2), I drew β_i^b, β_i^w from independent truncated bivariate normal densities, conditional on dependent mean vectors $\rho W \breve{\mathfrak{B}}^b$ and $\rho W \breve{\mathfrak{B}}^w$. Parameters $\breve{\mathfrak{B}}^b$ and $\breve{\mathfrak{B}}^w$ are scalars—vectors of values determined according to the EI extended model's functional form, or a preliminary set of draws from an independent bivariate truncated normal (W is a real-world spatial contiguity matrix based on all countries available, and ρ is a spatial autocorrelation scalar). To ensure that conclusions were not dependent on one set of parameters, the simulation was repeated over a wide range of values.

References

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

Donoho, D. L., and P. J. Huber. 1983. "The Notion of the Breakdown Point." In *A Feschrift for Erich L. Lehmann*, eds. Peter J. Bickel, Kjell A. Doksum, and J. L. Hodges. Belmont, CA: Wadsworth.

Imai, Kosuke, and Galy King. 2002. *Did Illegally Counted Overseas Absentee Ballots Decide the 2000 U.S. Presidential Election?* Cambridge, MA: Harvard University. http://gking.harvard.edu/preprints.shtml#ballots.

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

King, Gary. 2000. "Geography, Statistics, and Ecological Inference." *Annals of the Association of American Geographers* 90:601–606.