H-Net Reviews
in the Humanities & Social Sciences

Gary King. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data.* Princeton: Princeton University Press, 1997. xxii + 342 pp. $39.95 (paper), ISBN 978-0-691-01240-7; $95.00 (cloth), ISBN 978-0-691-01241-4.

Reviewed by James L. Huston (Oklahoma State University)

Published on H-Pol (March, 1998)

It's Time to Redo the Tables

Historians and political scientists in the United States are blessed with a cornucopia of election data from states, counties, townships, and precincts that covers nealy two hundred years. They are cursed because it is not certain what they should do with it. All this data takes the form of aggregate data, and, ever since William S. Robinson's influential article in 1950, scholars have been aware of the ecological inference problem: statistical results drawn from data sets in which voters have been aggregated into geographic units do not necessarily yield information about individual behavior. Most researchers want to generalize about individual voters, not their pattern of aggregation. Various techniques have been developed to overcome the aggregation problem, most notably Leo Goodman's regression solution for 2 by 2 tables, J. Morgan Kousser's system of "double regression" (the terminology that King uses) for larger tables, and W. Phillips Shively's methods of bounds. All have limitations, some of them severe. Until this book by Gary King, ecological inference was basically the equivalent of drinking dirty water: you never really knew the disease you were bound to contract. Gary King has forever altered our attitude toward ecological inference and its potential. He has provided an insightful excursion into the nature of the problem, the approaches researchers have applied, and the pitfalls fallen into. More to the point, he has constructed a model that remedies those ills and thereby allows individual behavior to be inferred from aggregate data. What elevates this study from others is that King has found data to test the validity of his model and produced diagnostics to determine when his model fails. Researchers in geography, economics, political science, and history are going to be forced to deal with King's model, and ultimately they will be greatly benefitted by it. At the risk of exaggeration, my guess is that King's approach to ecological inference may be the most important methodological discovery in political science during the past half-century. This is indeed an admirable, even an inspiring, book for its venture into theoretical reasoning, technical knowledge, interdisciplinary foundations, and practical forumulation. Please note, however, that I said the book is admirable; I did not necessarily mean that it is comprehensible.

The implications of this book need to be laid out and grasped in all their awesome significance. Every table that has used regression techniques to estimate individual behavior from aggregate data is wrong; the work of twenty-five years has just been invalidated. Though King does not exactly say this, his language leads to the conclusion that all cross-level inferences based on regression techniques are horribly flawed: "we know with absolute certainty that this 'constant parameter' assumption is incorrect" (p. 59). All the interpretations based on those tables–usually voter transition tables–are now suspect. King also advises us that the units of analysis best suited for ecological inference are the smallest available (i.e., precinct, township), are homogeneous rather than heterogeneous (does anyone remember the battles over this question in the 1970s?), and are plentiful, numbering in the hundreds or preferably thousands (this last point makes many studies based on counties in one state somewhat problematic).

Certainly King's discussion of the ecological inference problem, his criticism of past approaches, and his formulation of a model to provide answers would have by themselves attracted immense attention. But that is not where the raw power of this book comes from. Un-

like virtually all other writers on the subject, King tests his theory and validates its estimates of parameters. In four chapters, he shows how much error the Goodman regression technique produces in examples of voter registration by race, poverty status by gender, black literacy estimates in 1910, and voter registration in Kentucky. In these problems, King knew the actual parameter of interest and therefore could determine how well his model worked. The example of poverty by sex in South Carolina was especially revealing. The true values were 0.129 (i.e, 12.9 percent) for males and 0.177 (17.7 percent) for females; King's method for ecological inference model predicted estimates of .13 and .16. The Goodman regression technique yielded estimates of -.20 for males and .50 for females (p. 220). Furthermore, the regression technique has no real means of determining confidence intervals for the predicted value–the typical tools associated with regression are essentially useless for ecological inference. King's method, however, does have diagnostic tools that can aid in deciding how much faith can be placed in an estimate. How soon academics apply King's procedures to their research agendas is uncertain, but because court cases involving discrimination in electoral districts are a driving force for obtaining valid ecological inferences, a group of researchers will have no choice; the courts will soon insist upon King's method for making ecological inferences. (This may be the first time a Supreme Court ruling, *Thornburg v. Gingles*–in which the Court approved estimates from ecological regression for decisions in voter discrimination cases–has to be overturned due to an improvement in statistical methodology.)

The nature of the inference problem is key to understanding why regression techniques have so absymally failed. Most of the time, regression techniques do not focus on the parameter of interest, that is the B* (the notation I will here use) that represents the transition in each precinct (or unit of analysis). Most of the time researchers have been obtaining a weighted average, B**, that is not the true parameter. So that is one problem (pp. 31-33). The other is aggregation bias. If no aggregation bias existed there would be no problem; estimates could be easily made. But all problems with ecological inference, King asserts, are essentially the same, the problem of aggregation bias–the pattern of grouping individuals together produces means for the units of analysis that reflect the grouping, not individuals. The essential way the aggregation bias appears is that the parameters of interest, say B*1 and B*2, are correlated with each other over the units of analysis, and are correlated with one of the variables representing a known aggregate quantity (say, for example, the aggregate proportion of the Democratic vote in a county or precinct). Regression analysis cannot resolve this problem, and the correlations make the central assumption of ecological/Goodman regression wholly inadequate (the Goodman assumption being that the transition is constant over all precincts, that is, the parameter B* is constant in each unit of analysis). Trying to improve the regression model by adding information about voters will never rectify the problem of aggregation bias due to correlations between the parameters of interest (B*s) and their relationship to one of the variables of aggregation. Finally, most ecological inference problems are beset by enormous amounts of heteroskedasticity, also reflective of the aggregation bias problem. Researchers have failed to remove that heteroskedasticity.

Thus we come to King's proposed model. It is one of the most complicated that I have seen in the general literature, and for historians, I think, far exceeds anything they have run into. The model is highly interactive; the researcher is required to supply information at various parts to make the predictions accurate. King is explicit on this: all information is welcome because any clues can help reduce aggregation bias. Finally, in preparation for what is to come, the method does not rely on regression or any simple model of analysis. The reader deserves some peek at the model, but the reader should be cautioned that I am not sure I know what I'm talking about.

The central insight of King into the ecological inference problem is that the normal distribution fails to provide the distribution necessary to estimate the parameters of interest. King's method is based, it seems to me, on using the given data to create a particular distribution volume; the last step is to draw samples from that distribution volume, the mean of which stands as the best estimate for the parameter of interest–and from which may be obtained valid standard errors and confidence intervals that can be used to determine whether the estimate has significance or not. The running example King employs is a 2 by 2 table with two parameters of interest; to avoid a number of complications, my references will be to that 2 by 2 table. First, the researcher obtains the data and calculates the B**s of interest, because those are the only ones known and they have to act as initial proxies for the true parameters of interest, B*1 and B*2. However, the values are determined by finding their representation in a truncated normal distribution in which B**s are confined to a 0.00 to 1.00 range. Next, one employs tomography maps and scattercrosses to find loose relationships

between the B**s. Tomography plots are the ones associated with magnetic resonancing in hospital MRI scans. From this information, five parameters of interest are then determined: B**1, B**2, the variances of both, and the correlation between them. The parameters are then modified, weighted in a sense, to offset heteroskedasticity and aggregation bias. Indeed, it is in the initial truncation of the normal curve and then in the modification of the B**s that the aggregation bias is reduced. Then the values are placed into a maximum likelihood function that is also truncated to implement the method of bounds advocated by Shively. The process King develops calls for three distinct reparameterizations (none of which I really understand). The last reparameterization was virtually impossible, and it is here that King uses a "monte carlo" or sampling approach to obtain finally estimates of the parameters of interest–or what historians typically call the estimates of transition. If anyone really understands what I have just written, they are further along the learning curve than I am.

Learning the method will be the great problem associated with King's book. Reviewers frequently write that a book was twice as long as it needed to be; I can honestly say this book is one-half the size it needs to be. And the difficulty is not language. Indeed, King writes better than most statisticians I have read; the book sparkles with wit, exuberance, and patient explanation, partially, I think, because King recognizes the importance of his discovery. Nonetheless, the mathematical background needed to understand the concepts is quite high, and in other places the ideas and representations are so novel that extensive discussion should have been indulged in. For instance, King introduces a series of graphical constructs that are indisputably unique and end up being insufficiently explained, especially if they are to be used as diagonistic tools and as means of estimating relationships between the B**s. King remarked in a discussion of tomography plots that "in order to help interpret the plots, which take some getting use to..." (p. 126). I would nominate that phrase for an understatement-of-the-year award. The graphs needed to employ this method require far more elaboration.

In most mathematical texts, the common practice seems to be that when a theorem is given (or, by analogy, a model), the practical application in terms of a problem is often two or three times the length of the theorem. In King's presentation, the model is about three times the length of the illustrations. Given the complexity of the model, it would have greatly aided comprehension for those who are statistically-impaired to have had a patient step-by-step demonstration of how the model informs an application. However, I expect those problems to be cleared up rather soon. King has a computer program available via the internet that maps out his method, although the most powerful application comes with a program called GAUSS, which is evidently popular among statisticians but fairly rare among historians.

I note, however, that in his acknowledgements King thanks Sidney Verba for his assistance; I therefore have little doubt that SPSS will soon be offering a capsule based on King's model and probably some extensive documentation as well. One additional feature should be mentioned, however: King's method will probably enormously increase the time required to create a transition table compared to the older method of ecological regression, probably by a ratio of 5 or 10 to 1. It remains to be seen whether King's method will yield satisfying results in tables larger than 2 by 2. Almost all transition tables historians develop are larger than 2 by 2, especially now that the political division is at least two parties and the category of nonvoters. The complications arising from correlations among the B**s ought to increase significantly and I would think estimates of what those correlations are might become difficult to discover. Yet my sense is that King's method will, perhaps after some alterations, finally win out. It may take several years before any results start appearing, for the gestation period for the methodology is going to be longer than it ever was for multiple regression.

King's extensive statistical formulations will have to receive verfication from sources other than myself (although I was puzzled in places about his calculations for the Goodman technique and his explanation of double regression). He has a proof to show that all problems connected with ecological inference are different perspectives of aggregation bias. Methodologists and theorists will have to pass judgment on those claims.

I imagine that for a number of individuals in political science, sociology, statistics, and geography, King's work may in fact be quite comprehensible and even forward. But my intuition is that few in the field of history are going to be able to absorb this book. It appears that statistical methodology in the other disciplines has become increasingly sophisticated, whereas history almost lacks a methodological branch. That disjuncture is worrisome because it means obvious breakthroughs like this one may be beyond the profession's ken.

This problem, the lack of a field in history devoted to statistical methodology, probably merits some consider-

ation and discussion.

Copyright (c) 1998 by H-Net, all rights reserved. This work may be copied for non-profit educational use if proper credit is given to the author and the list. For other permission, please contact H-Net@h-net.msu.edu.

If there is additional discussion of this review, you may access it through the network, at:

https://networks.h-net.org/h-pol

Citation: James L. Huston. Review of King, Gary, *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data.* H-Pol, H-Net Reviews. March, 1998.

URL: http://www.h-net.org/reviews/showrev.php?id=1792

Copyright © 1998 by H-Net, all rights reserved. H-Net permits the redistribution and reprinting of this work for nonprofit, educational purposes, with full and accurate attribution to the author, web location, date of publication, originating list, and H-Net: Humanities & Social Sciences Online. For any other proposed use, contact the Reviews editorial staff at hbooks@mail.h-net.msu.edu.