# The Ecological Fallacy

Paul A. Jargowsky
University of Texas at Dallas



**GLOSSARY**

**aggregation** – in this context, the process of grouping data on individual units of analysis, such as calculating neighborhood-level means of individual-level variables.

**bias** – the difference between the expected value of an estimator and the value of the parameter being estimated.

**correlation** – the degree of linear association between two variables.

**ecological fallacy** – the assumption that relationships between variables at the aggregate level imply the same relationships at the individual level.

**ecological inference** – a conclusion about associations or causal relationships among individual observations at one level based on the analysis of data aggregated to a higher level.

**omitted variable bias** – bias in a estimator resulting from the omission of a relevant variable when the omitted variable is correlated with one or more of the explanatory variables.

In many important areas of social science research, data on individuals is summarized at higher levels of aggregation. For example, data on voters may be published only at the precinct level. The ecological fallacy refers to the incorrect assumption that relationships between variables observed at the aggregated, or ecological level, are necessarily the same at the individual level. In fact, estimates of causal effects from aggregate data can be wrong both in magnitude and direction. An understanding of the causes of these differences can help researchers avoid drawing erroneous conclusions from ecological data.

## I. Origins of the Ecological Fallacy

The ecological fallacy has a long history spanning many disciplines, particularly sociology and political science. It is closely related to what economists tend to call aggregation bias. Stated briefly, *one commits an ecological fallacy if one assumes that relationships observed at an aggregated level imply that the same relationships exist at the individual level.* For example, observing that the percent black and the crime rate are correlated at the level of police precincts does not necessarily imply that blacks are more likely to commit crimes. Indeed, it is possible that the correlation of two variables at the aggregate level can have the opposite sign as the correlation at the individual level. As a result, it can be quite difficult to infer individual-level relationships from aggregated cross-sectional data, an issue known as the problem of ecological inference.

Social science is mostly about understanding the behavior of individuals. Quite often, however, researchers find that the only data available to address certain empirical questions is aggregate data. For example, a researcher may wish to know whether a specific racial group is more inclined to vote for a particular party. However, since balloting is secret, the researcher does not have access to the individual-level data. Instead, he or she may know the vote total for two parties in each election precinct and the demographic characteristics of the voting age population in each of these precincts. *Ecological inference* refers to this process of attempting to draw an inference about individual relationships from aggregate data. In this example, the researcher would attempt to draw a conclusion about the relationship between the race of an individual and his or her voting propensity from the relationship between two precinct-level

variables: the precinct proportion in the racial group, and the precinct proportion voting for a particular party. More generally, ecological inference is the drawing of a conclusion about how X affects Y in some population of interest from data which consists of the means of X and Y for subgroups of the population.

A. Robinson's Critique of Ecological Inference.

Although not the first to draw attention to the problems of ecological inference, Robinson (1950) had the most dramatic impact. Robinson cited a number of famous studies from several disciplines that were based on what he called "ecological correlations" (351-352). That is, the cited studies relied on ordinary Pearsonian correlation coefficients between two variables calculated from the *averages* of those variables for spatially-defined groups of individuals, such as neighborhoods, cities, states, or regions. These studies had assumed, often implicitly, that the implications that could be drawn from the sign, magnitude, and significance of the ecological correlations applied equally to the relationship between the two variables at the level of individuals, which in almost all cases was the primary objective of the research.

Robinson subjected this practice to a withering critique, by contrasting individual and ecological correlations in cases where data were available at both levels. He showed that the individual-level correlation between race and illiteracy in the U.S. in 1930 was 0.203, but the correlation between percent black and percent illiterate at the state level was far higher, 0.773. Robinson showed that not even the sign of ecological correlations could be trusted. The correlation between having been born abroad and being illiterate was a positive 0.118 at the individual level (again using 1930 data for the U.S.), probably reflecting the lower educational

standards of the immigrants' countries of origin. However, the correlation at the state level between the corresponding ecological aggregates (percent foreign born and percent illiterate) was a counterintuitive -0.526, the opposite direction of the individual correlation! Robinson concluded that "there need be no correspondence between the individual correlation and the ecological correlation" (354). Moreover, he said he provided "a definite answer as to whether ecological correlations can validly be used as substitutes for individual correlations." His answer: "They cannot" (357).

The impact of Robinson's condemnation of ecological inference was profound. Indeed, inferences about individual relationships from aggregate data came to be regarded not just as problematic, but – though Robinson's seminal article did not use this word – as a fallacy (Selvin 1958: 615). And while Robinson's critique was stated in terms of simple bivariate correlation coefficients, his critique is a challenge regression analysis on aggregate data as well. All slope coefficients in bivariate and multiple regressions can be expressed as functions of either simple or partial correlation coefficients, respectively, scaled by the standard deviations of the dependent and independent variables. Because standard deviations are always positive, the sign of any regression coefficient reflects the sign of the correlation coefficient on which it is based, whether simple or partial. Thus, regression analysis on aggregated data – a common practice in several disciplines – runs the risk of committing the ecological fallacy as well.

B. The Limitations of Robinson's Critique.

Seen in retrospect, Robinson's analysis seems to ignore the presence of confounding variables. For example, using Robinson's second example, immigrants tended to flock to

industrial states in search of jobs, and these states are wealthier and had higher literacy rates than poor (jobless) Southern states that failed to attract as many immigrants. To a modern reader, Robinson's analysis seems to lack appropriate controls for socioeconomic status, regional dummy variables, or a fixed-effects model to isolate the effect of illiteracy from other covariates. Indeed, Hanushek et al. revisited Robinson's data and showed that the sign of his correlation was a reflection of left-out variable bias (1974: 90-95); in other words, his model was underspecified.

If the anomalous results attributed to the ecological fallacy actually result from model mis-specification, then "the ecological fallacy itself is a near fallacy" (Firebaugh 1978: 570). On the other hand, if the divergence between individual and aggregate level estimates are more subtle and intractable, then ecological inference is a dangerous business. The following section illustrates the mathematical bases of the ecological fallacy, which in turn gives some guidance as to how it can be avoided.

## II. Understanding the Mathematical Structure of the Ecological Fallacy

To understand the ecological fallacy, one needs to understand what causes the differences between estimators generated by data at difference levels. The next section provides graphical illustrations that establish how ecological inference can go wrong. Sections B and C develop the two mathematical conditions that cause such aggregate estimates of relationships between variables to differ from their individual-level counterparts.

### A. Graphical Illustration of the Problem with Ecological Correlations.

We begin by considering a few simplified scenarios using scatterplots, following Gove

and Hughes (1980). Suppose we are interested in a dichotomous dependent variable such as dropping out of high-school, coded as either 1 if a person is a dropout or 0 if the person is not. Further, suppose there are two groups, white and black, and the basic question of interest is whether members of one group or the other are more likely to drop out. But we lack data on individuals. Instead, we only know the overall proportion of persons who are dropouts in 3 different neighborhoods. We also know the proportion black in each of the three neighborhoods, which for the purpose of illustration I have set to 0.20, 0.50, and 0.80.

Figure 1 shows how ecological inference is supposed to work. The figure shows the separate rates for whites and blacks as dashed lines, because the researcher does not observe these data. The black group has a higher dropout rate than the white group, and so as the proportion black in the neighborhood rises the overall dropout rate also rises. In this case, one could correctly infer from the aggregate data that blacks are more likely to drop out.

[Figure 1 about here.]

Figure 2 shows how the ecological data can give misleading results. In this case, whites have a higher dropout rate than blacks in each neighborhood. However, the dropout rate of both groups rises as percent black in the neighborhood rises, perhaps because percent black in the neighborhood is correlated with some other variable such as family income. Even though whites have higher rates than blacks in every neighborhood, the ecological regression coefficient will have a positive slope, because the overall dropout rate rises as percent black rises. In this case, the ecological regression would correctly report that the DV is positively associated with percent black in the neighborhood, but the inference that individual blacks are more likely to drop out than whites would be wrong. (This is an example of what is known as Simpson's paradox.)

[Figure 2 about here.]

Other scenarios are possible. Suppose the black and white dropout rates are exactly the same within each neighborhood, but the rates for both groups rise as the percent black in the neighborhood rises. At the ecological level, the observed dropout rates will slope upwards, even though there is no effect of race at the individual level. Figure 3 shows a case where the dropout rate rises as the percent black rises solely because the *whites* have higher rates in the neighborhoods in which they are the minority. Again, an inference from the ecological level that blacks drop out more often would be incorrect.

[Figure 3 about here.]

In Figure 4, blacks do have higher drop out rates than whites in each neighborhood, and the rates of both groups rise as percent black increases. Regression on the aggregate data produces a positive slope, but virtually all of that slope is driven by the common increase of both groups in the more heavily minority neighborhoods. Only a small fraction of the slope reflects the influence the race of individuals on the drop out rate. In this case, the direction of the ecological inference would be correct, but the magnitude of the effect would be substantially overestimated.

## B. How Aggregation Produces Bias

The problem described above can be restated as a form of aggregation bias (Freedman 2001; Irwin and Lichtman 1976; Stoker 1993; Theil 1955). We want to understand how one variable affects another in the population. In other words, we want to know the slope parameter that tells us how the dependent variable changes in response to changes in the independent

variable. We can obtain an estimate of the effect by applying ordinary least squares (OLS) to the following regression equation:

$$Y_{ij} = \alpha + \beta X_{ij} + u_{ij}, \qquad [1]$$

in which *j* indexes neighborhoods and *i* indexes individuals within neighborhood *j*. The OLS estimate of the slope, *b*, has the following expected value:

$$E\left[b \mid X\right] = \beta + E\left[\frac{\sum_j \sum_i X_{ij} u_{ij}}{\sum_j \sum_i \left(X_{ji} - \bar{X}\right)^2}\right]. \qquad [2]$$

This result is analogous to the standard proof that OLS coefficients are unbiased, found in any econometrics textbook, except for the double summation sign. However, in view of the associative property of addition, the double summation signs does not affect the sums or the conclusion. If the second term in equation 2 is zero, which will occur if and only if *X* and *u* are uncorrelated, then *b* is an unbiased estimate of β. The disturbance term, however, implicitly includes the effect of all other variables as well as random influences. If the net effect of these omitted variables and influences is correlated with *X*, the assumption is violated, the second term does not reduce to zero, and the estimator is biased.

By summing up to the neighborhood level and dividing by the number of observations in each neighborhood, equation 1 implies that:

$$\bar{Y}_j = \alpha + \beta \bar{X}_j + \bar{u}_j. \qquad [3]$$

The β that appears in equation 3 is algebraically identically to the β in equation 1. Thus, in principle, an estimate of the effect can be obtained from either the individual or the aggregate level regressions. However, the expected value of the slope estimate from the aggregate

regression, $b^*$, is:

$$E\left[b^* \mid X\right] = \beta + E\left[\frac{\sum \bar{X}_j \bar{u}_j}{\sum \left(\bar{X}_j - \bar{X}\right)^2}\right]. \qquad [4]$$

At the aggregate level, the condition for the unbiasedness is that the *mean* disturbance term is not correlated with the *mean* value of the independent variable. It is quite possible that in a given set of data, the criteria for unbiasedness is met at the individual level (equation 2), but violated at the aggregate level (equation 4). Such a correlation could arise if the grouping process is related to some variable Z, not included in the regression, which is correlated with the outcome variable (Freedman 2001). Thus, the ecological fallacy can arise from a particular kind of left out variable bias, one that is introduced or exacerbated because of the aggregation process. In addition, a correlation between $X$ and $u$ at the aggregate level could arise if the grouping process is based on the values of the dependent variable, $Y$; in that case, either extreme values of $X$ or extreme values of $u$ would produce extreme values of $Y$, that would then tend to be grouped together.

Although we have explicated these ideas in the context of a bivariate regression, they apply equally well in the multiple regression context. In fact, in the world of social phenomena, where there are always correlations among explanatory variables, it is highly unlikely to be the case that a bivariate regression would be correctly specified at either the individual or the aggregate level. In the multivariate context, however, one additional problem arises. An outcome variable for an individual may be affected by both the individual's value of X and by a contextual variable that is a function of the aggregated values of X, such as the mean of X (Firebaugh 2001: 4025). When the individual data are aggregated, the individual and contextual

values of X may not be separately identified. This problem is discussed further in Section C below.

The first implication of the foregoing discussion is that if both the individual and ecological regressions are correctly specified, both types of analyses will provide equally unbiased estimates of the true slope parameter. In symbolic terms,

$$E[b \mid X] = E[b^* \mid X] = \beta \qquad [5]$$

The second implication is that both regressions can be mis-specified, and in the later case there is no guarantee that the individual regression is the better of the two. Grunfeld and Griliches (1960), referring to individual regressions as micro equations and ecological regressions as macro equations, argue that ecological regressions may be better in certain circumstances:

> [I]n practice we do not know enough abut micro behavior to be able to specify micro equations perfectly. Hence empirically estimated micro relations...should not be assumed to be perfectly specified.... Aggregation of economic variables can, and in fact frequently does, reduce these specification errors. Hence, aggregation does not only produce aggregation error, but may also produce an aggregation gain. (1)

It is not hard to think of examples where aggregation could reduce correlation between the disturbance term and X. For example, persons may choose their neighborhoods on the basis of unobserved characteristics which also affect their wages. In that case, neighborhood characteristics will be correlated with the disturbance term in a wage regression, resulting in biased estimates of the neighborhood effect on wage. Aggregating to the metropolitan level would sharply reduce this source of bias, by subsuming all neighborhood-to-neighborhood

selection in the metropolitan averages.

The third implication is that is possible to think about the conditions under which the bias term in equation [4] has an expectation different from zero. Assume that we can write down a well-specified individual model based on individual-level variables, as in equation 1, but only lack the data to estimate it. If the same equation estimated at the aggregate level produces biased estimates, then there must be something about the grouping mechanism that leads to correlation between the relevant X variables and the disturbance term. In other words, it matters how the data were aggregated. It is useful to consider the following possibilities and their implications:

1. Random grouping is not very likely to arise in practice, but it is instructive to consider the possibility. If the data are aggregated randomly, and the model was correctly specified at the individual level, there will be no aggregation bias. The expected value of mean *X* and mean *u* for all groups will be the grand mean of *X* and *u* respectively, and they will not be correlated.

2. If the grouping is based on the *X* (or multiple *X*s), there will be no aggregation bias. This follows because the conditional mean of the disturbance term is zero for all values of *X* if the individual model is correctly specified.

3. If the grouping is based on *Y*, aggregation bias is very likely. For example, if *Y* and *X* are positively related, in the groups with higher levels of *Y* one would find both high values of *X* and larger than average disturbance terms, and at lower levels of *Y*, the opposite would occur. Clearly, the aggregate levels of *X* and *u* will be correlated and the ecological regression is mis-specified.

4. Grouping based on geography, the most common method, is also the most difficult to evaluate, since neighborhood selection may be based on a complex set of factors operating at different levels. However, if the dependent variable is something like income, the danger exists that neighborhood aggregation is more like case 3. If the dependent variable is less likely to be involved in the residential choice function, then sorting by neighborhood will be more like cases 1 or 2.

When data are provided in an aggregate form, the researcher must understand and

evaluate how the groups were formed. Then the researcher must try to ascertain whether the procedure is likely to introduce aggregation biases or aggregation gains in view of the specific dependent variable and explanatory models under consideration.

C. Problems Related to Group-Level Effects.

The forgoing discussion is based on equation 1 and, like most empirical literature in social science, this equation does not take into account the possibility of group level effects on individuals. That is, the individual level equation only includes group level variables. But it is possible, indeed likely, that the mean value of X in a neighborhood could have an independent effect on Y even after controlling for the individual's own level of X. Firebaugh (1980) describes several possibilities. An intelligent student may well learn more in the presence of more intelligent fellow students. On the other hand, a mediocre student might be discouraged in such an environment and do better if he was "a big fish in a small pond." Group effects include or are related to neighborhood effects, peer group effects, and social network effects.

In general, we can characterize these models as including some measure of a group level variable in an individual model:

$$Y_{ij} = \beta_1 + \beta_2 X_{ij} + \beta_3 \bar{X}_j + u_{ij} \qquad [6]$$

At the aggregate level, this model becomes:

$$\begin{aligned} \bar{Y}_j &= \beta_1 + \beta_2 \bar{X}_j + \beta_3 \bar{X}_j + \bar{u}_j \\ &= \beta_1 + (\beta_2 + \beta_3)\bar{X}_j + \bar{u}_j \end{aligned} \qquad [7]$$

Clearly, even if there is no bias of the type discussed in the previous section, the individual and

group effects are not identified, only their sum. In the absence of individual data or outside information on the magnitude of one or the other of the two effects, the lack of identification in the aggregate models poses a formidable obstacle.

Fortunately, in certain cases, the sum of the two effects may itself be of interest. For example, suppose the dependent variable is a measure of children's health, and X is a measure of insurance coverage through a public program. One might expect a direct impact of the child's own coverage status, as well as an effect of the level of coverage in his or her area, through reduction of contagious diseases and increased availability of medical service providers (a supply response). Both effects are real benefits of the program, and both are included in the coefficient from the ecological regression.

## III. Solutions to the Ecological Inference Problem?

While Robinson's critique sent shock waves through the social science community and undoubtedly influenced some researchers to eschew aggregate data, it also spawned a literature on "solutions" to the ecological inference problem. Goodman (1953, 1959) addressed the problem in terms of dichotomous variables. He noted that the dependent variable at the aggregate level is a proportion, which must be the weighted sum of the unobserved proportions of the two groups formed by the independent variable. This is just an accounting identity. In the case of voting, we observe the overall proportion voting for a given party and wish to make inferences about the votes for specific individuals depending on their racial group. The weighted average of the two groups' voting must add to the observed total proportion in each neighborhood:

$$T_i = (1 - P_i)W_i + P_i B_i \quad [8]$$

where $T_i$ is the observed proportion, $P_i$ is the percent black, and $W_i$ and $B_i$ are the unobserved rates for the white and black sub-populations, respectively.

Algebraic manipulation yields an equation which can be estimated from the aggregate data:

$$T_i = W_i - (W_i + B_i) P_i = \alpha + \beta P_i + u_i \quad [9]$$

The constant term the regression is average proportion voting for the party in the white population, and $\beta$-$\alpha$ produces the estimate of the black proportion. The disturbance term is introduced because $\alpha$ and $\beta$ are fixed, whereas in actuality $W_i$ and $B_i$ vary from neighborhood to neighborhood. The validity of this approach depends on the "constancy assumption"; in other words, the voting proportions do not depend on the ethnic composition of the neighborhood (Goodman 1953, 1959; Freedman 2001). Figure 1 illustrated a case of the constancy assumption, because the white and black drop out rates were unrelated to the percent black.

A second basic approach is based on establishing bounds for the minimum and maximum possible for each cell of a cross-tabulation in each of the aggregate units (Duncan and Davis 1953). By summing these extrema up over the data set, it is possible to determine with 100 percent confidence the minimum and maximum bounds of the correlation that could obtain in the individual level data.

King (1997) proposed a "solution" to the ecological inference problem, dubbed "EI." It was also developed in the context of dichotomous dependent variables. EI combines the method of bounds with Goodman regression technique, and estimates the system using maximum

likelihood and numerical simulation, assuming a bivariate normal distribution for the parameters. Critics have pointed out a number of flaws with King's technique, a review of which are beyond the scope of this essay. Important critiques are Anselin (2000), Anselin and Cho (2002), Freedman (1998), and McCue (2001).

The debate on the statistical underpinnings and empirical performance of the EI method will likely continue for some time, even as the technique is being widely adopted within the field of political science. However, the most important issue concerning King's approach is that it is developed within and justified for a very narrow range of problems that are not fully representative of the range of issues and types of data historically associated with the ecological fallacy and the problem of ecological inference. King dismisses the argument that ecological inference is mainly a matter of model specification, and in doing so reveals the most serious problem in his proposed methodology. "[T]he concept of a 'correctly-specified' individual-level equation is not helpful in this context," he argues,

> since individual data contain the answer in ecological inference problems with certainty. That is, with individual data, we would not need to specify any equation; we would merely construct the cross-tabulation and read off the answer. Having the extra variables around if indivdual-level data are available would provide no additional assistance. (49)

In other words, the narrow focus of King's technique is reconstructing a description of the individual data, not evaluating a causal model. This is an adequate goal in King's motivating example, ascertaining voting patterns by race for the purpose of redistricting litigation. But in virtually any other social science application, our interest is in a causal model that can not be reduced to a contingency table. Even in voting analysis, there are substantively interesting

questions about whether racial identity affects voting net of other factors, such as income, occupation, and so on. Further, King readily acknowledges that his method will be less effective when the dependent variable is continuous, because no information is gleaned from bounds (p. 260). These are rather important limitations.

For further discussion of approaches reduce bias in ecological inference, see Achen and Shively (1995), Cho (2001), and Freedman (1991, 2002).

## IV. Conclusion, or Learning to Live With Aggregate Data

Anselin, in a review of King's work, put it best: "There is no solution to the ecological inference problem" (2000: 589). There are only estimates based on assumptions, but this is also true about regressions on individual-level data. No single procedure can claim to be the solution to the ecological inference problem. In the absence of data about individuals, one can derive estimates about individual relations only by carefully specifying a model, and these assumptions must be guided by theory, experience, and consistency with observable relations.

When social scientists attempt to analyze aggregate data, the best course of action is to parameterize the variables relevant to the grouping process as well as possible. As noted previously, Hanushek et al. (1974) were able to show that the real problem with Robinson's data was an underspecified model, not aggregation. For example, it would be particularly important to control for race and income if the data are neighborhood aggregates. If there are contextual effects, these need to be modeled as well, perhaps using multi-level models (Brown and Saks 1980; Firebaugh 1978: 570). Firebaugh (1999: 4025) proposes adding additional independent variables which explain and hence control for contextual effects.

One can be misled by ecological correlations or by regressions on aggregate data, but one can be equally misled by simple correlations or regressions based on individual data, and for some of the same reasons – left out variables, model mis-specification, and false assumptions about the process under study. Robinson's 1950 article generated five decades of productive debate over the ecological fallacy and related topics such as ecological inference, aggregation bias, and contextual effects. With multivariate analysis, advanced modeling techniques, and an understanding of the aggregation process, researchers can mostly avoid falling victim to the ecological fallacy. Indeed, in certain specific situations, aggregate data may be better than individual data for testing hypotheses, even if those hypotheses are about individual behavior. The "ecological fallacy" has lost some of its sting, and should not cause researchers to abandon aggregate data.

## Further Reading


Achen, C. H. And W. P. Shively (1995). *Cross-Level Inference.* Chicago: University of Chicago Press.

Anselin, Luc (2000). "The Alchemy of Statistics, or Creating Data Where No Data Exist." Book Review. *Annals of the Association of American Geographers* 90:586-92.

Anselin, Luc and Wendy K. Tam Cho (2002). "Spatial Effects and Ecological Inference." *Political Analysis* (forthcoming).

Brown, Byron W. and Daniel H. Saks (1980). "Economic Grouping and Data Disaggregtion." In *New Directions for Methodology of Social and Behavioral Science*, vol. 6, *Issues in Aggregation*, edited by K. Roberts and L. Burstein. San Francisco, California: Jossey-Bass.

Cho, Wendy K. Tam (2001). "Latent Groups and Cross-Level Inferences." *Electoral Studies* 20: 243-263.

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

Firebaugh, Glenn (1978). "A Rule for Inferring Individual-Level Relationships from Aggregate Data." *American Sociological Review* 43:557-72.

——— (1980). "Groups as Contexts and Frog Ponds." In *New Directions for Methodology of Social and Behavioral Science*, vol. 6, *Issues in Aggregation*, edited by K. Roberts and L. Burstein. San Francisco, California: Jossey-Bass.

_____ (2001). "Ecological Fallacy, Statistics of." Pp. 4023-4026 in Neil J. Smelser and Pual B. Baltes, eds., *International Encyclopedia of the Social and Behavioral Sciences..* Oxford, UK: Elsevier.

Freedman, D.A. (2001). "Ecological Inference." Pp. 4027-4030 in Neil J. Smelser and Pual B. Baltes, eds., *International Encyclopedia of the Social and Behavioral Sciences..* Oxford, UK: Elsevier.

_____, S.P. Klein, M. Ostland, and M.R. Roberts (1998). "A Solution to the Ecological Inference Problem." Book Review. *Journal of the American Statistical Assocation* 93:1518-22.

_____, S.P. Klein, J. Sacks, C.A. Smyth, C.G. Everett (1991). "Ecological Regression and Voting Rights." *Evaluation Review* 15: 673-711.

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

_____ (1959). "Some Alternatives to Ecological Correlation." *American Journal of Sociology* 64:610-25.

Gove, Walter R. and Michael Hughes (1980). "Reexamining the Ecological Fallacy: A Study in Which Aggregate Data Are Critical in Investigating the Pathological Effects of Living Alone." *Social Forces* 58:1157-77.

Grunfeld, Yehuda and Zvi Griliches (1960). "Is Aggregation Necessarily Bad?" *Review of Economics and Statistics* 42:1-13.

Hanushek, Eric, J. Jackson, and J. Kain (1974). "Model Specification, Use of Aggregate Data, and the Ecological Correlation Fallacy." *Political Methodology* 1: 89-107.

Irwin, Laura and Allan J. Lichtman (1976). "Across the Great Divide: Inferring Individual Level Behavior from Aggregate Data." *Political Methodology* 3:411-39.

King, Gary (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton, New Jersey: Princeton University Press.
McCue, Kenneth F. (2001). "The Statistical Foundations of the EI Method." *The American Statistician* 55:106-10.
Selvin, Hanan C. (1958). "Durkheim's Suicide and Problems of Empirical Research." *American Journal of Sociology* 63: 607-619.
Stoker, Thomas M. (1993). "Empirical Approaches to the Problem of Aggregation Over Individuals." *Journal of Economic Literature* XXXI:1827-74.
Theil, Henri (1955). *Linear Aggregation of Economic Relations*. Amsterdam: North-Holland.

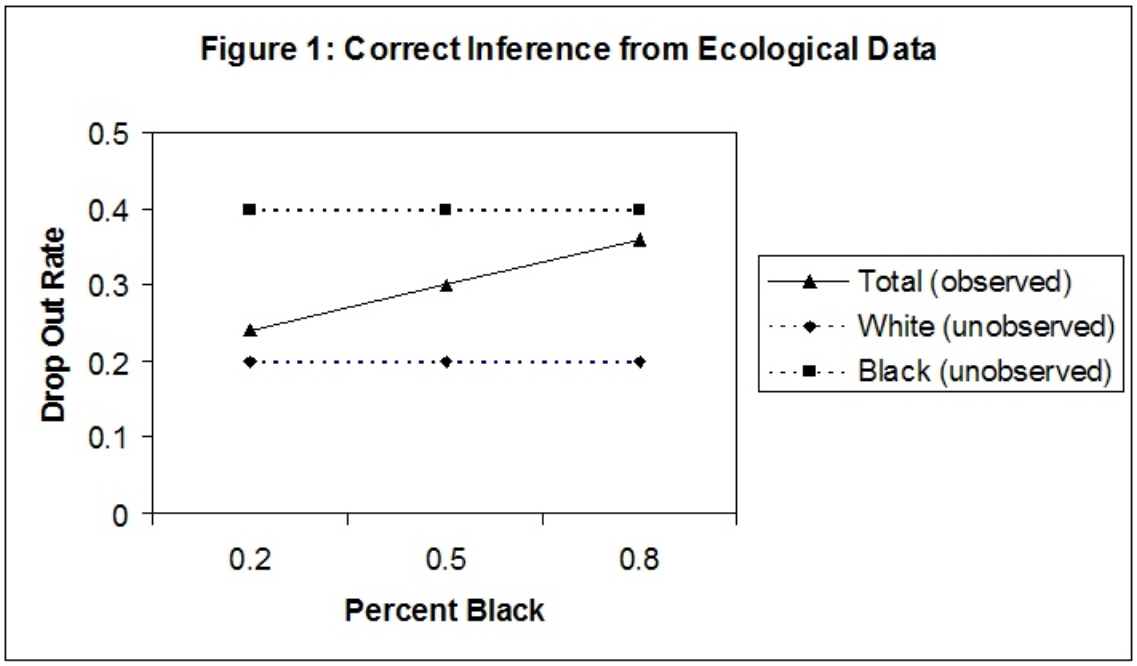
Figure 1: Correct Inference from Ecological Data
0.5
0.4
0.3
0.2
0.1
0
Drop Out Rate
0.2
0.5
0.8
Percent Black
Total (observed)
White (unobserved)
Black (unobserved)

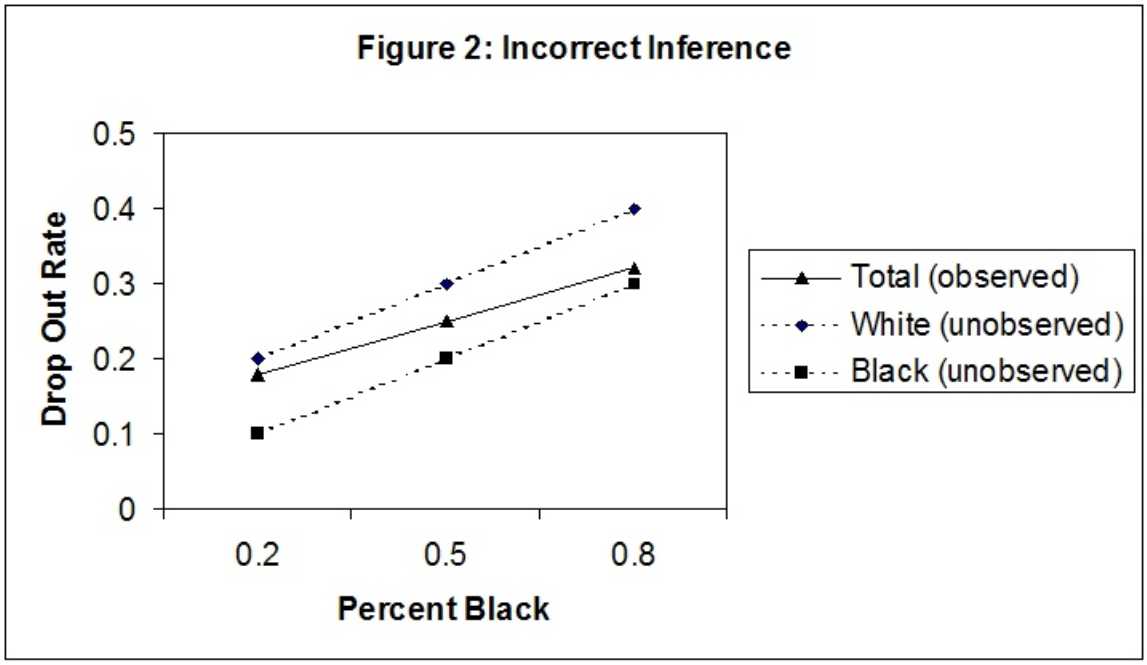
Figure 2: Incorrect Inference
0.5
0.4
0.3
0.2
0.1
0
Drop Out Rate
0.2
0.5
0.8
Percent Black
Total (observed)
White (unobserved)
Black (unobserved)

**Figure 3: Incorrect Inference**

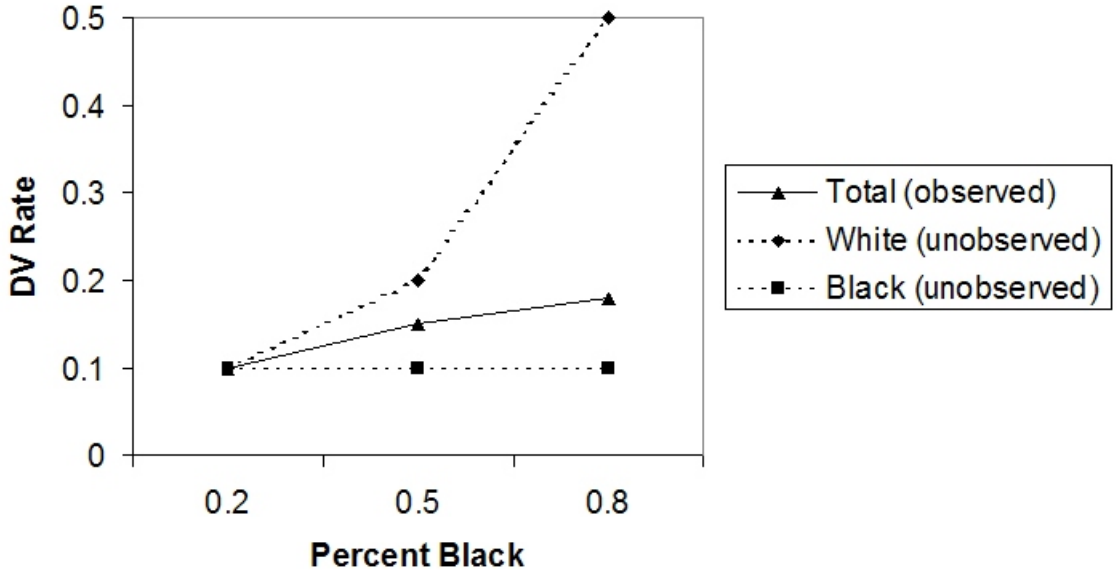


**Figure 4: Correct Inference, Wrong Reason**