Quantitative Geographical Analysis

Week 4: The Ecological Problem.

Background paper for February 12, 2001 seminar discussion.

Elvin Wyly

1. Oikos Logos.

The word “ecology” occupies a curious position in the social and natural sciences. At first glance, it’s obvious: the term usually refers to the study of “the interactions of organisms with each other, with their biotic and abiotic environments, and the outcomes of these interactions.”\(^1\) Yet the lexicon has become a battleground in the last generation, as influential scholars have explored the epistemological boundaries between an objective, ‘scientific’ biological ecology and a “normative ecology” concerned with human constructions of the idea of ‘nature,’ social and political institutions governing human-nature relations, and so forth.\(^2\)

Perhaps such ambiguity should come as no surprise, given the origins of the word. In 1866, the German biologist Ernst Haeckel coined the term ‘oecologie,’ from the Greek word oikonomia. This delicious morsel is the root of ‘economy,’ variants of which were used in Greek, then Latin, and then French to describe the principles and practices of household management.\(^3\) Ecology can also be translated more directly from the Greek oikos (house) and logos (knowledge/word/discourse). We have, therefore, two meanings: the ‘science’ of the household, and the ‘management’ of the household. What Haeckel had in mind was a strong metaphor: “nature’s economy.”\(^4\)

The metaphor of nature’s economy had deep and lasting impacts on thought across many disciplines in the early years of the twentieth century. One prominent example was sociology, where influential scholars portrayed the seeming chaos and complexity of a nascent urban industrial modernity (especially in the large cities of the United States) in terms of metaphors drawn from ecology and biology.\(^5\) A field of “human ecology” emerged in the 1920s at the

---


\(^2\) See McManus (2000) for a concise and valuable summary of the various movements, including social ecology, political ecology, global ecology, and several others.

\(^3\) This etymology lends a particularly vicious tone to Lloyd Rodwin’s comment, relayed by Wilson, and then by Fotheringham (1993) in his reflective essay: “Putting the word ‘urban’ before ‘economics’ has the same effect as putting the word ‘domestic’ before ‘science.’” Contemporary criticism is one thing, but when you can marshall the discursive authority of Latin or Greek, you’re really shaking that mythical ivory tower. Fotheringham, A.S. (1993). “On the Future of Spatial Analysis: The Role of GIS.” Environment and Planning A, Anniversary Issue, 30-34.


\(^5\) The specific metaphors deployed to understand urban phenomena included community, competition, disturbance, equilibrium, and the notion of “invasion and succession” as new social groups (i.e. species) colonized and then dominated a new neighborhood (i.e., environment). The human ecology literature shares common ground with, but is arguably distinct from, the lineage of the Chicago School of Sociology. But the overarching theme of this work is either social-Darwinist or resolutely status-quo: the observed social and spatial order of the urban industrial world is that which is required to ensure its success and reproduction.
University of Chicago, and motivated a broad tradition of inquiry on social conflict and
competition in the context of an intricate environment produced by economic development,
industrialization, and urbanization. Simultaneously, sociology and the other social sciences were
engaging with new developments in inferential statistics (which really took off in the 1930s), and
were exploiting the growing availability of quantitative data. Typically, however, investigators had
to make do with aggregate or summary data for populations in specific geographic areas — what
came to be called ‘ecological’ data — when they were really interested in behaviors and
relationships at the individual level.

So this etymological-intellectual tour brings us to the “ecological fallacy.”

2. The Problem.

In 1950, W.S. Robinson published “Ecological Correlations and the Behavior of
Individuals” in the American Sociological Review.\(^6\) His concern was with the increasingly prevalent
use of areal summary data to infer individual-level relationships and processes. He demonstrated
the risks of ecological inference by performing a series of correlation analyses.

A Short Detour to \(r\)

So we need to recall a bit of the details of correlation analysis. Remember that the
correlation coefficient is a measure of the strength of association between two variables measured
on an interval scale.\(^7\) The coefficient is normally denoted as \(r\), and depends on the degree to
which the variance in one variable coincides with the variance in another. From last week’s notes,
recall that variance is nothing more than a measure of ‘spread,’ calculated as the sum of the
squared deviations from the mean, divided by the number of observations:

\[
\frac{\sum (x^2)}{N}
\]

Where little \(x\) is the difference between each observation and the variable mean. So we can express
variance as:

\[
\frac{\sum (X - \mu_x)(X - \mu_x)}{N}
\]

If we have two variables, we can measure their association by making one small change to the
variance equation. We then have the covariance:

\[
\frac{\sum (Y - \mu_y)(X - \mu_x)}{N}
\]

Where Y is our second variable of interest. The covariance, of course, is still in the units of the original variables; in order to make it possible to compare things measured in different units, we need to adjust the covariance with the product of the standard deviations of each of the variables:

$$r = \frac{\text{Covariance}}{\sigma_x \sigma_y}$$

When we write out the full equations for the covariance and standard deviations, and get rid of those pesky, redundant terms that cancel themselves out in the numerator and denominator, we have this:

$$r = \frac{\sum (Y - \mu_Y)(X - \mu_X)}{\sqrt{\sum (Y - \mu_Y)^2 \sum (X - \mu_X)^2}}$$

The correlation coefficient always ranges between -1 and +1. Values close to -1 indicate a strong inverse relationship (as one variable increases, the other decreases); values close to +1 indicate a strong direct relationship (as one variable rises, the other does too). Values near zero signify no relationship. Finally, when we square the correlation coefficient, we obtain the coefficient of determination, or $r^2$. The coefficient of determination measures the proportion of the variance in one variable that is accounted for by variance in another variable.

**Robinson’s findings**

Robinson analyzed data from the U.S. Census of 1930. He focused on the relationship between race and literacy, calculating $r$ for (Y) the percentage of each state’s population over age 10 who were illiterate, and (X) the percentage of each state’s population over age 10 who were Black. The resulting coefficient was 0.773, indicating a strong direct relationship. Performing the analysis at the more spatially aggregated level of the nine Census regions increased the coefficient to a robust 0.946, indicating an unusually strong relationship. At the individual level, however, this link all but disappeared: for the 97 million persons over age 10 in the nation at the time, the correlation was 0.203. An analyst using spatially aggregated data would conclude that there was a strong relation between race and illiteracy, when in reality there was less than a 5 percent chance ($0.203^2$) that a black person was also illiterate. Robinson performed the same analysis comparing illiteracy to the foreign-born population, and found a sign reversal: aggregate data at the state level yielded a strong inverse correlation ($r=0.619$), while at the individual level $r$ was 0.118.

Robinson’s work inaugurated widespread attention to the risks of inferring individual relationships from aggregated data. One of those who spearheaded a major research agenda on measurement issues and ecological problems was the sociologist Hubert Blalock. Blalock performed a series of experiments to test various hypotheses on the effects of aggregation on bivariate relationships. In addition to his concern for the sensitivity of $r$ – the strength of

---

7 The correlation coefficient can also be calculated with variables coded as dummy indicators. Ordinal data can be analyzed with Spearman’s rank correlation coefficient.
association of variables - Blalock was also wanted to know if aggregation bias affected the form of a relation.

**Recalling Regredi**

So at this point we need to recall a few of the specifics of regression.\(^8\) For present purposes, we just need a brief crash course in a few of the highlights; if this doesn’t look familiar, you should consult a few basic statistics texts to see how different authors present alternative routes to the same set of concepts.\(^9\) Consider a simple, two-variable dataset with just ten observations:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>58</td>
</tr>
<tr>
<td>43</td>
<td>73</td>
</tr>
<tr>
<td>18</td>
<td>56</td>
</tr>
<tr>
<td>10</td>
<td>47</td>
</tr>
<tr>
<td>25</td>
<td>58</td>
</tr>
<tr>
<td>33</td>
<td>54</td>
</tr>
<tr>
<td>27</td>
<td>45</td>
</tr>
<tr>
<td>17</td>
<td>32</td>
</tr>
<tr>
<td>30</td>
<td>68</td>
</tr>
<tr>
<td>47</td>
<td>69</td>
</tr>
</tbody>
</table>

With the elementary formulas of last week’s background paper, we obtain the mean and standard deviation for X and Y:

\[
\mu_x = 29 \\
\sigma_x = 11.42 \\
\mu_y = 56 \\
\sigma_y = 11.80
\]

Using the formula on the previous page, we obtain an \(r\) of 0.72, indicating that 52 percent of the variation in \(Y\) can be accounted for by variation in \(X\). When we graph the relationship between \(X\) and \(Y\), we obtain the scatterplot shown on the following page. The purpose of regression is to fit a straight line that describes the form of the relationship between two (or more) variables, such that

---

\(^8\) Why is it called regression? The word comes from the Latin *regredi*, “to go back,” and was used by nineteenth-century researchers to describe reversion to the mean. Francis Galton, in studies of the heredity of height and other physical characteristics, observed that very tall people tend to have children somewhat shorter (closer to the mean) than themselves; short people tended to have children a bit taller than themselves. Regression towards the mean influences the degree to which knowing one variable helps to predict another: with very little regression, prediction can be quite accurate; with a great deal of regression to the mean, we can only predict poorly if at all.

changes in a dependent variable (Y) can be predicted by changes in an independent variable (X). The predictions in Y can be denoted by:

\[ \hat{Y} \]

and since this is only an estimate, the real value of Y will be a function of the regression estimate and an error term:

\[ Y = \hat{Y} + e \]

The predicted values of Y are expressed as a straight line, for which all we need are an intercept term (a) and the product of a slope coefficient (b) and the independent variable:

\[ \hat{Y} = a + bX \]

Obtaining the most accurate predictions of Y requires that we minimize the sum of the error terms; we could simply sketch in a variety of different lines, and add up the deviations between the observed values of Y and the predictions; unfortunately, each of these lines would, paradoxically, yield error sums of zero – because the overpredictions would cancel out the underpredictions. Working around this problem is simple: we square the error terms, and find a solution that minimizes the sum of squared deviations. This is “ordinary least squares,” or OLS regression.

Differentiation yields the solution to the problem of minimizing the sum of squared errors. The sum is at a minimum when:
\[ b = \frac{\sum XY - \sum X \sum Y}{\sum X^2 - (\sum X)^2 / N} \]

and once we crack this annoying little beast of sigmas, tampering with the why equals ay plus bee ex equation at the mean gives us:

\[ a = \bar{Y} - b\bar{X} \]

There are several different ways of rewriting that annoying beta equation, and one of the more friendlier versions is this:

\[ b = \frac{\sum xy}{\sum x^2} \]

Where little \( x \) and little \( y \) are the deviations between each observation and its respective mean. So we take our table of \( X-Y \) values and work through the necessary calculations:

<table>
<thead>
<tr>
<th>( X )</th>
<th>( x )</th>
<th>( x^2 )</th>
<th>( Y )</th>
<th>( y )</th>
<th>( xy )</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>11</td>
<td>121</td>
<td>58</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>43</td>
<td>14</td>
<td>196</td>
<td>73</td>
<td>17</td>
<td>238</td>
</tr>
<tr>
<td>18</td>
<td>-11</td>
<td>121</td>
<td>56</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>-19</td>
<td>361</td>
<td>47</td>
<td>-9</td>
<td>171</td>
</tr>
<tr>
<td>25</td>
<td>-4</td>
<td>16</td>
<td>58</td>
<td>2</td>
<td>-8</td>
</tr>
<tr>
<td>33</td>
<td>4</td>
<td>16</td>
<td>54</td>
<td>-2</td>
<td>-8</td>
</tr>
<tr>
<td>27</td>
<td>-2</td>
<td>4</td>
<td>45</td>
<td>-11</td>
<td>22</td>
</tr>
<tr>
<td>17</td>
<td>-12</td>
<td>144</td>
<td>32</td>
<td>-24</td>
<td>288</td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>1</td>
<td>68</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>47</td>
<td>18</td>
<td>324</td>
<td>69</td>
<td>13</td>
<td>234</td>
</tr>
</tbody>
</table>

\[ \sum xy = 971 \quad \sum x^2 = 1304 \]

Giving us a beta coefficient of 0.7446. Then \( a=56(.7446)(29) \), or 34.4066. We can now use the fitted regression line to express both the form and the strength of the relation: a one-unit increase in \( X \) leads to a corresponding increase of 0.7446 units in \( Y \), and this effect accounts for 52 percent \( (r=0.72^2) \) of the variation in \( Y \). There are additional nuances to bivariate regression, which involve tests for the possibility that \( r \) could have occurred by chance (critical values of pearson’s \( r \), tests to measure the dispersion of points around the regression line (the ‘standard error of the estimate’), and so forth. The situation evolves into delicious levels of detail in the multivariate case, even
though the basic idea remains the same. But for the moment, $r$ and $b$ are sufficient to grasp the essence of Blalock’s work on aggregation.

*The Blalockian Conjecture*\(^\text{10}\)

Blalock observed that when individual observations were averaged across areal zones, and when small zones were grouped into larger ones, the effect was to reduce the variance in the dependent variable; idiosyncratic factors that are important at the individual or local level tended to cancel one another out with progressive aggregation. Correlation coefficients, therefore, grew larger at higher levels of aggregation. Interpreting this result, however, was no simple matter: on the one hand, we could conclude that changes in scale imply changes in the relationship between the variables. But it may also mean that at one scale the relationship is hidden by other, unmeasured factors. Specifically, bivariate correlation and regression abstract from the innumerable influences among different variables, some of which may be quite important. If we are interested in only two variables ($x$ and $y$), but ignore other crucial factors (say, $i$, $j$, and $k$), we may have something like this:

$$
\begin{align*}
& \text{i} \\
\downarrow & \\
& \text{j} \\
& \downarrow \\
& \text{y} \\
\downarrow & \\
& \text{x}
\end{align*}
$$

What happens to the x-y link with aggregation depends in large part how grouping affects the omitted variables i, j, and k. Blalock performed a simple series of tests to illustrate this problem. He began by measuring median income for whites and blacks in 1950 for 150 counties in the American south, calculating a racial disparity for each county. He compared this gap ($y$) to the percentage of each county’s population that was black ($x$), obtaining an $r$ of 0.54, with a beta coefficient of 0.26.

\(^{10}\) I might be a bit strange, but as Dave Barry might say, I am not making this up. Stan Openshaw and Peter Taylor, in a sweeping overview of the modifiable areal unit problem (MAUP) and alternative solutions, conclude: “An alternative approach towards a statistical theory is based on a Blalockian conjecture (see Taylor, 1977, p. 222). This fails because it completely ignores the existence of aggregation effects and because the basic idea is far too simple for such a horribly complex problem.” William of Occam might disagree. See Openshaw, S., and Taylor, P.J. (1981). “The Modifiable Areal Unit Problem.” In N. Wrigley and R.J. Bennett, eds., *Quantitative Geography: A British View*, 60-69. London: Routledge and Kegan Paul.
Blalock then grouped the 150 counties into progressively larger groups, according to three simple rules: 1) purely random grouping, 2) maximizing the variance in x, and 3) by spatial proximity. The first and the third rules are self-explanatory; the second entails ranking the counties by their x values, and then grouping counties based on their rank in this list (i.e., grouping the first- and second-ranked, the third- and fourth-ranked, and so on for pairs).

This is what Blalock found:

<table>
<thead>
<tr>
<th>Ungrouped</th>
<th>Pairs</th>
<th>Fives</th>
<th>Tens</th>
<th>Fifteens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_{xy}$</td>
<td>$b_{xy}$</td>
<td>$r_{xy}$</td>
<td>$b_{xy}$</td>
<td>$r_{xy}$</td>
</tr>
<tr>
<td>Random</td>
<td>.54</td>
<td>.26</td>
<td>.67</td>
<td>.36</td>
</tr>
<tr>
<td>Max Var(x)</td>
<td>.54</td>
<td>.26</td>
<td>.67</td>
<td>.26</td>
</tr>
<tr>
<td>Proximity</td>
<td>.54</td>
<td>.26</td>
<td>.63</td>
<td>.27</td>
</tr>
</tbody>
</table>

Consider the logic of each rule. With random grouping, we expect that the variance in x and y declines in tandem, and that the same thing happens to the unmeasured, omitted variables i, j, and k. Thus we would expect to see random fluctuations in the correlation and beta coefficients, with no systematic pattern. The results confirm this expectation, although the coefficients fluctuate rather wildly when we aggregate into ten groups of fifteen (r drops to 0.26, b slips to 0.18).

The second experiment is more interesting. When we group the counties in such a way as to preserve the maximum variance in x, the effect is to introduce systematic bias into only one part of a three-component system. The effect of grouping on the other two – y, and the unmeasured variables i, j, and k, which are unrelated to x – is random. As a consequence, Blalock hypothesized that the variance in y and i, j, and k would decline more rapidly than that in x. The latter would account for a rising share of the variance in the dependent variable (y). His experiments confirmed this result, and also revealed that the form of the relationship – the beta coefficient – remained stable. Systematic aggregation increases the strength of an observed relationship, but does not affect its functional form.

Finally, the third experiment presents us with the usual dilemma of geographical research, where units are grouped according to spatial proximity. In this case, the a priori hypothesis depends upon the degree to which grouping by spatial proximity will systematically affect the variance of x. If x is highly spatially autocorrelated, then the result will be a progressively stronger relationship between x and y. Blalock’s experiments confirmed this finding for racial disparities in income and the spatial concentration of African Americans in poor counties of the “Black Belt” of the Southern Piedmont. At mid-Century, the deep historical legacy was still stark in Blalock’s analysis: formal and de facto segregation kept African Americans confined to poor, mostly rural counties where class divisions remained sharp and rigid.
A Flock of Fallacies?

Blalock’s work provided important insights. We now know that the effects of aggregation are not completely random, and that it is generally much safer to rely on estimates of the form of a relationship rather than its strength. Nevertheless, variations of this kind of problem have plagued quantitative applications for half a century. Several recurrent themes have been important. First, one variant of the ecological problem remains intractable – the “modifiable areal unit problem” or MAUP. The essence of this problem is that simply changing the boundaries of areal units will affect the results; using the Blalock example, the problem is that the entire analysis rests on the use of counties, and there is no guarantee that these are appropriate for the phenomenon under investigation. Indeed, in most applications the use of a particular zonal definition is dictated by data availability, not any coherent rationale regarding the phenomena under investigation. The effects can be substantial. In what remains the landmark experimental work on the MAUP, Stan Openshaw decomposed the ecological problem into two components: one was a scale effect, which was essentially the same as in Blalock’s experiments; the other was an aggregation effect, which captured the bias introduced by differences in how boundaries were among areal zones. Openshaw showed that there was an almost infinite number of ways to aggregate individual-level data to zonal definitions, and that the (bivariate) correlations generated by alternative zonal configurations could be understood as a frequency distribution. Although the distribution was tightly clustered around a mean, the possible values of r ranged almost the full range from -1 to +1, leading Openshaw to title one chapter “A Million or So Correlation Coefficients.”

The second problem involves a broader set of considerations. To the degree that the ecological problem is defined in narrow, technical terms, analysts have devised ways to deal with it (the readings for this week offer a new solution that builds on and refines tools developed not long after Robinson’s article was written). But the broader point – the risks of inference across different scales of process and analysis – cannot be so readily dispatched. Over the years, researchers have recognized a number of common fallacies, and it is rare for any piece of geographical research not to succumb to at least one. Ron Johnston provides a valuable compendium:

1. the individualist fallacy: assuming that the whole is no more than the sum of its constituent parts. Processes generating strong correlation coefficients at the county level, for example, cannot simply be aggregated up to understand the operation of state-level processes.

2. the cross-level fallacy: assuming that a relationship observed in one grouping holds in all other combinations. This is essentially the same as the MAUP.

---


3. the universal fallacy: assuming that relations observed in one selection of individuals holds for the population. The validity of this assumption depends upon whether the selection was defined according to the principles of random sampling.

4. the selective fallacy: the use of carefully-chosen cases to provide evidence for one interpretation.

5. the cross-sectional fallacy: assuming that relations observed in one time period hold at other times.

This is by no means a recipe for analytical paralysis. Scientific (and humanistic) understanding must advance, research results must be shared with colleagues, and policies must be evaluated or enacted. No piece of research is perfect, and none ever will be. But this typology of fallacies provides a useful ‘gold standard’ by which we can evaluate results, whether they be preliminary output on the computer screen, or tables in a published journal article or book. It also serves to keep us sensitive to the implicit assumptions of our work, and helps to remind us of the considerations and caveats that should be written into our own research.

3. A Solution.

Gary King, a political scientist at Harvard, has recently proposed an innovative solution to the ecological inference problem. Refining earlier attempts to work around the ecological problem, King has devised a way of inferring certain types of ecological relations. King begins with a problem that is common in electoral studies - and that, in light of last year’s encounter with dimpled chads, dueling lawsuits and press conferences, and allegations of racial intimidation at the polls, is rather timely. Suppose we know:

\[ X_i \] the proportion of the voting-age population in zone i that is African American, and

\[ T_i \] the proportion of the voting-age population casting ballots (the overall turnout in each precinct).

(For the sake of simplicity, assume that we can divide the population solely into two, mutually-exclusive categories – black and non-black.) The problem is that we do not have information on the black (and non-black) turnout for each precinct. Is it possible to infer the race-specific turnout for the entire area under study, or for individual precincts?

Goodman (1953) proposed a simple solution.\(^\text{13}\) Consider what we get if we graph \( X_i \), the racial composition variable, against \( T_i \), overall turnout:

On average, more or less, all other things being equal, what we have is simple: among more African-American precincts (X), voter turnout (T) is generally lower. This dilemma is the basis of longstanding attempts by the NAACP, Jackson’s Rainbow Coalition, and scores of other groups to get out the black vote. For the statistical purposes of King, however, note the similarity to the simple regression tutorial above. If we regress T on X, the Beta coefficient gives us an overall estimate of the Black turnout, while r would give us some estimate of the confidence that could be placed in the inference. This is shown in the red line through the scatter of points above.

Goodman’s solution stood as the best approach for a number of years, but was far from perfect - it often yielded unrealistic results (e.g., 120 percent or -2 percent turnout), and did not allow for different rates of turnout for different groups in different precincts. King essentially re-worked the approach, and replaced Goodman’s single regression with a model that allows Beta coefficients to vary across different precincts. The graph then appears like that shown on the following page.
Our goal this week is to understand a) how King derives these estimates, b) what they mean and how the approach can be used, and c) its contributions and limitations. The work is not simple, and the jargon can be thick at times. But work through it carefully, and try to tease out the essential points. Read Fotheringham’s review of King’s book closely, and, if you want to retrace the steps I will be following for next week, track down some of the citations in the review, and go to


where King has made some of his papers ~ and the software that generates the output shown above ~ available for download.

Sift through the readings for this week, taken from a special book review issue of the Annals of the Association of American Geographers, and tell me what you think.


