INADVERTENT SOCIAL THEORY: AGGREGATION AND ITS EFFECT ON COMMUNITY RESEARCH

By Peter H. Greenwood and A. E. Luloff

Although continued attention has been given to the general study of "community," we still lack a consensus, operational definition. Such an absence impedes development in this area of study. Because authors have used different areal conceptualizations, knowledge is, at best, case specific. This paper reviews the problem most often associated with aggregation of data, namely, aggregation bias, and considers some additional problems not usually associated with aggregation. The review serves as a prelude to a discussion of community research methodology which, at present, is beset with potential aggregation problems. Interpretation and possible implications of these findings are advanced.

INTRODUCTION

The purpose of this paper is to review and extend the understanding of the potential consequences of working with aggregated data. We argue that not only may the use of aggregated data lead to biased results, but also that the use of aggregated data leads to unreliable tests of hypotheses. The significance of these findings will be developed within the context of community research.

Unaggregated data is data which is collected for the smallest possible unit of analysis for the shortest period of time. In all practical instances, therefore, a researcher deals with aggregated data; however, for the purpose of this paper unaggregated data is data which is collected for a unit of analysis appropriate to the research interest and data collected for other units of analysis are either disaggregated or aggregated.

Social scientists have long known about the bias which can occur with the use of aggregated data. In his seminal paper Robinson discussed a phenomenon which he labelled the ecological fallacy. In this study Robinson demonstrated that correlations among variables are influenced by the unit of analysis. Since then a number of authors have elaborated on his findings and issued similar warnings. Despite these warnings, social scientists continue to employ data which has been aggregated to a level which differs from the level of their interests.

From time to time a researcher may face the dilemma of using aggregated data or using no data at all. The traditional counsel has been that the researcher must choose between potentially biased estimators or none at all. Although Firebaugh has presented a rule to determine if a problem with bias exists, his rule is of little utility to our researcher. It is possible, however, that the situation is not as bleak as the traditional counsel suggests.

Suppose that the researcher's interest lies in the effect of some variable x on some variable y for a specific unit of analysis, and that he hypothesizes that a simple linear equation describes the relationship between the variables. For example, he may hypothesize that:

$$(1)\ y_{ij} = \alpha + \beta x_{ij} + \varepsilon_{ij}$$

where y_{ij}, x_{ij}, and ε_{ij} are the values of y, x, and a disturbance for the j'th unit in the i'th group, and α and β are unknown constants. The values of these constants may be estimated with ordinary least squares (OLS): to perform the calculations, observations on y and x for a sample of n units should be assembled. These n equations, each of the form given by (1), may be written compactly as:

$$(2)\ Y = Xb + \varepsilon$$

where Y and ε are column vectors of the y observations and the disturbances, X is a two column matrix where the first column is a unit vector and the second contains the x observations, and b is a column vector of the unknown constants. The OLS estimates of b are given by the well-known formula:

$$(3)\ \hat{b} = (X'X)^{-1} X'Y.$$

Substituting (2) into (3) leaves:

$$(4)\ \hat{b} = b + (X'X)^{-1} X'\varepsilon.$$

If, as is normally assumed, x_{ij} and ε_{ij} are uncorrelated and if the expectation of ε_{ij} is zero, $\hat{b}$ provides unbiased estimates of b; that is, the expectation of $\hat{b}$ is b. Normally each x_{ij} is assumed to be a fixed number, and each disturbance is determined independently of any value in X. The researcher's misfortune is that such a sample is unavailable in this case.

The b vector can also be estimated with observations which represent aggregations of micro-data. The prior estimates were derived from observations at the unit level; it is also possible to derive estimates from observations made at the group level. Perhaps the most meaningful comparison is provided by simply aggregating the earlier sample. The sample can be aggregated quite easily; Feige and Watts simply premultiply both sides of (2) by a conversion matrix d. In this case (2) becomes:

$$(5)\ dY = dXb + d\varepsilon.$$

The vector dY now contains aggregations of the observations in Y; these aggregations may be simply the sum of the y's in each group, or they may be the average of the y's. Premultiplying by (dX)' leaves:

$$(6)\ (dX)'dY = (dX)'dXb + (dX)'d\varepsilon$$

$$\text{or } X'd'dY = X'd'dXb + X'd'd\varepsilon$$

$$\text{or } X'AY = X'AX + X'A\varepsilon$$

where A is d'd.

The OLS estimators of b derived from the aggregated sample are given by:

$$(7)\ \hat{b}_{ag} = (X'AX)^{-1} X'AY$$

There is no reason to expect that $\hat{b}_{ag}$ will correspond to $\hat{b}$ even though the same X and Y are used in their derivations. This does not mean that $\hat{b}_{ag}$ is biased; the question of bias remains. Substituting from (5) leaves:

$$(8)\ \hat{b}_{ag} = b + (X'AX)^{-1} X'A\varepsilon.$$

Feige and Watts argue that the expectation of $(X'AX)^{-1} X'A\varepsilon$ depends on the aggregation rule employed even if the expectation $(X'x)^{-1} X'\varepsilon$ is zero. This is obviously true. Aggregation bias

Peter H. Greenwood is Assistant Professor of Resource Economics and A. E. Luloff is Assistant Professor of Community Development, University of New Hampshire. Published with the approval of the Director of the New Hampshire Agricultural Experiment Station as Scientific Contribution No. 917.

is present whenever the aggregation rule is related to the disturbances.

The conclusion of Feige and Watts and Firebaugh would appear to indicate that the dilemma of our researcher is irresolvable. However, this dilemma may not be as serious as it appears. The easiest way to see this is to expand the compact notation and investigate the formulas in greater detail. For example, the estimator of β implied by (4) is:

$$(9)\ \hat{\beta} = \beta + \frac{n\Sigma x_{ij}\varepsilon_{ij} - \Sigma x_{ij}\Sigma\varepsilon_{ij}}{n\Sigma x_{ij}^2 - \Sigma x_{ij}\Sigma x_{ij}}.$$

By assumption each x_{ij} is a fixed number; that is, $E(x_{ij}\varepsilon_{mk}) = x_{ij}E(\varepsilon_{mk})$, where E is the expectations operator and ε_{mk} is simply any one of the error terms. Thus, if $E\varepsilon_{ij} = 0$, then $E\hat{\beta} = \beta$, and $\hat{\beta}$ is an unbiased estimator of β.

If the sample were aggregated by calculating the group means, the β estimator implied by (8) is:

$$(10)\ \hat{\beta}_{ag} = \beta + \frac{N\Sigma X_i e_i - \Sigma X_i \Sigma e_i}{\Sigma X_i^2 - \Sigma X_i \Sigma X_i};$$

$X_i = \Sigma_j x_{ij}$ where n_i is the number of units in the i'th group. Each n_i

e_i would be computed in a similar fashion from the disturbances, and N is simply the number of groups. Whether or not $\hat{\beta}_{ag}$ is biased depends on the expectation of the fraction residuum in (10). If the fraction is further expanded, the numerator is seen to contain many terms; each term contains the product of an observation of x and an observation of ε. Furthermore, each of these terms is weighted by the product of the relevant n's. For example, one term in the numerator is

$$(11)\ \frac{Nx_{ij}\varepsilon_{ij}}{n_i^2}$$

The expectation of this term is:

$$(12)\ Nx_{ij}E\left[\frac{\varepsilon_{ij}}{n_i^2}\right]$$

since each x is fixed. If each x is fixed and if $E\varepsilon_{ij}$ is zero, an unaggregated sample yields unbiased estimators. For $\hat{\beta}_{ag}$ to also be unbiased it is only necessary to assert that $E(\varepsilon_{ij}n_i) = 0$. In practice this condition is not restrictive. Bias would be present if small groups (low n) had relatively large errors or vice versa. The problem in terms of β can be eliminated simply by aggregating with sums rather than means in any event. While one may rig an example such that $E(\varepsilon_{ij}n_i) \neq 0$, it is our opinion that in practice no sleep should be lost worrying about this potential source of bias. To recapitulate: if the model is correctly specified and amenable to unbiased estimation with unaggregated data, it is usually the case that the same model is amenable to unbiased estimation with aggregate data. Moreover, since in any case where n_i is error correlated, the model of (1) is incorrectly specified; therefore, the term "usually" in the previous sentence could as well be replaced with "always". If the researcher has confidence in the hypothesized model, the problem of aggregation does not lie with potential bias of the estimators. To this extent the researcher's dilemma dissolves, but another problem of aggregation, which has not received the attention that bias has, remains.

The remainder of this section considers the impact of aggregation on routine statistical measures. Assume that (1) is a correct model, and for the sake of concreteness suppose that the sample consists of T counties which can be transformed into a sample of m regions. It will simplify the algebra if we further suppose that each region contains n counties (nm = T). We first consider the impact of aggregation on the t statistics.

The T observations enable us to estimate α, β, and the variances of their estimators σ_α^2, σ_β^2. An unbiased estimator of $\sigma_{\alpha\beta}^2$ is:

$$(13)\ \hat{\sigma}_{\alpha\beta}^2 = \frac{\Sigma_i\Sigma_j(\hat{y}_{ij} - y_{ij})^2}{(T-2)\Sigma_i\Sigma_j(x_{ij} - \bar{x})^2} \qquad \begin{matrix} i = 1, \ldots, m \\ j = 1, \ldots, n \end{matrix}$$

where $\hat{y}_{ij} = \hat{\alpha} + \hat{\beta}x_{ij}$, and $\bar{x} = \frac{\Sigma_i\Sigma_j x_{ij}}{T}$ • Simplify with $\Sigma\Sigma(\hat{y}_{ij} - y_{ij})^2 = \underset{r\ r}{\Sigma\varepsilon^2}$,

$\Sigma\Sigma(x_{ij} - \bar{x})^2 = \Sigma\Sigma\Delta x^2 ij = \underset{r\ r}{\Sigma\Delta x^2}$. To test whether $\hat{\beta}$ is significantly different from zero, we calculate the t ratio:

$$(14)\ t = \hat{\beta}\left[\frac{(T-2)\Sigma\Delta x^2}{\Sigma\varepsilon^2}\right]^{1/2}$$

which we recall is distributed as a students' t with (T-2) degrees of freedom.

Had we used, instead, the regional means to estimate α and β ($\hat{\alpha}'$ and $\hat{\beta}'$ to distinguish them from the earlier estimates), the t ratio calculated would be:

$$(15)\ t' = \hat{\beta}'\left[\frac{(m-2)\,(\Sigma\Delta x^2 + \underset{ij\quad k}{\Sigma\Sigma}\Delta x_{ij}\Sigma\Delta x_{ik})}{\Sigma\varepsilon'^2 + \underset{ij\quad k}{\Sigma\Sigma}\varepsilon'_{ij}\Sigma\varepsilon'_{ik}}\right]^{1/2}, \text{ where } j \neq k.$$

The term, $\Sigma_i\left[\underset{j\quad kik}{\Sigma\Delta x_{ij}\Sigma\Delta x}\right]$, appears in the numerator; a similar term with the residuals replacing the Δx's appears in the denominator. This term instructs us to compute the sum of all the Δx cross products within a region for each region, and sum the results.

Comparing t and t' we find:

$$(16)\ \frac{t}{t'} = \frac{\hat{\beta}}{\hat{\beta}'}\left[\frac{T-2}{m-2}\,\frac{\Sigma\Delta x^2(\Sigma\varepsilon'^2 + \Sigma\Sigma\varepsilon'\Sigma\varepsilon')}{(\Sigma\Delta x^2 + \Sigma\Sigma\Delta x\Sigma\Delta x)\Sigma\varepsilon^2}\right]^{1/2}.$$

If the estimators, $\hat{\alpha}'$ and $\hat{\beta}'$, are unbiased any correlation among the residuals within regions is coincidental, and thus $\Sigma\Sigma\varepsilon'\Sigma\varepsilon'$ should be small. However, even in the event that identical estimates were obtained and $\Sigma\Sigma\varepsilon'\Sigma\varepsilon'$ were zero we would still be left with:

$$(17)\ \frac{t}{t'} = \left[\frac{T-2}{m-2}\,\frac{\Sigma\Delta x^2}{\Sigma\Delta x^2 + \Sigma\Sigma\Delta x\Sigma\Delta x}\right]^{1/2}.$$

Even under these favorable conditions we would be unable to determine whether t would be larger than t'. If regions have similar counties, t' will tend to be greater than t; this implies that a hypothesis we might reject at the county level could be accepted at the regional level. Similarly, there are hypotheses that we would reject at the regional level that we would accept at the county level.

R^2 estimates are also affected by the level of partial aggregation chosen by the researcher. The T county observations would generate an R^2 given by:

$$(18)\ R^2 = 1 - \frac{\Sigma\varepsilon_r^2}{\Sigma\Delta y_r^2}.$$

The value of R^2 obtainable with regional means may be approximated by asking what percentage of the regional mean y's is explained by $\hat{\alpha}$ and $\hat{\beta}$, which we recall were obtained from the county data. The approximation must be lower than the true value. We calculate:

$$(19)\quad r^2 = 1 - \frac{\Sigma\varepsilon'^2}{\Sigma\Delta\bar{y}^2}$$

Comparing R^2 and r^2 we find

$$(20)\quad r^2 - R^2 = \frac{\Sigma\varepsilon^2}{\Sigma\Delta y^2} - \frac{\Sigma\varepsilon^2 + \Sigma\Sigma\varepsilon\Sigma\varepsilon}{\Sigma\Delta y + \Sigma\Sigma\Delta y\Sigma\Delta y}$$

When comparing the t statistics, the sum of the residual cross products was assumed to be small; however, the Δy cross products need not be small, since if x is similar within a region, y will also be similar. Because r^2 is an approximation, (2) understates the difference in R^2 when regions have similar counties. R^2 tends to increase as one aggregates if one aggregates similar units.

DISCUSSION

In this section the consequences of aggregation will be discussed and developed in the context of a problem area for empirical research—the study of communities. Two good reasons for this selection are: (1) there is an expanding interest in community research, and (2) the existence of potential aggregation problems is easily demonstrated. Community research is not, however, unique in suffering from the consequences of aggregation.

Problems of measurement frequently influence the selection of key variables and/or the unit of analysis employed in social science research. Community research offers no exception to this generalization. For example, a number of studies have attempted to determine whether or not there are significant differences between urban and rural communities; however, measurement in some has been at the individual level, and in others the county has been the unit of analysis. These studies may claim to address the issue, even when recognizing that neither the individual nor the county is the ideal unit. The selection of county data is frequently justified by the following line of reasoning:

(1) The county is the one administrative unit below the level of the state for which the greatest amount of comparative data is available;

(2) The use of city data alone eliminates the rural population and would prohibit the measurement of the effect of the urban-rural determinants within the community system. Furthermore, even if some more precise "locality" designation would be preferable (city, town, village etc.) comparative data are readily available only for cities larger than 25,000; and

(3) The political, social, economic, cultural and functional boundaries of cities and villages are more sharply delineated than are those of counties (Bonjean, Browning, and Carter, p. 150).

These arguments notwithstanding, we know that if the concept of a community is not coterminous with county lines then the results derived from county observations are misleading since the results may or may not be supported solely because of the effect of aggregation.

Social scientists will continue to employ aggregated data; in some cases the researcher has no choice. For instance, aggregation is sometimes necessary to protect the confidentiality of respondents. In other cases researchers rely on published data which are available only in an aggregated form. In such cases caution must be exercised when the results are interpreted. Additionally, aggregation may occur by default as when researchers cannot agree on the level of aggregation which most clearly corresponds with a theoretical construct.

We place community literature in the latter case. Many researchers have selected their data sets on the basis of vastness, completeness, or comparability, but in so doing they have laid a groundwork for inadvertent social theory and policy. It is reasonable to assume that interest in areas such as education, religion, and local community services may develop in smaller political units than the county. Similarly, studies of school consolidation revealed that citizens often report the importance of maintaining some form of local identification at the neighborhood level. In recent studies of achievement (Bidwell and Kasarda, 1975; 1976; Hannan, et al.; Alexander and Griffin), the importance of the politically recognized subdivision of the school district as the unit of analysis has led to a discussion of the problem of specification of research models. Evidence further suggests that interactional fields can exist within many larger areas (Kaufman; Kaufman and Wilkinson).

Similarly, the argument of Bonjean, Browning, and Carter indicates the preferability of using local units as the focus of inquiry. Their contention, even if true, that political, social, economic, cultural and functional boundaries are no more sharply delineated for counties than for villages or cities does not prohibit the study of smaller units. Additionally, the claim that there is not enough comparable data on smaller political units seems unwarranted. Although it is true that it takes more time to compile, since such data are rarely collated for small units, frequently *the data are available.* Finding the source(s) presents a problem, but once located the benefits may outweigh the investments in time and energy.

Further, the fact that municipalities are smaller areal units than metropolitan areas or counties contributes to greater specificity in measurement of structural features. When counties are used, it is not clear whether or not a service institution is readily and equally accessible to all of the residents of a county, since multiple political subdivisions are included within the county. The literature on location of health services well establishes this point (cf. Ball and Wilson; Breisch; Coleman; Scheffler).

Finally, concern with local decision making is growing at the federal level. Governments have all too often based their policy analysis on social science investigations of large political units. While the success or failure of many public policies depends on the response of cities or counties to federally mandated programs, the continued existence and, indeed, growth, of many smaller places indicates a need to study community responses as well.

While it is true that aggregation bias is not usually present when a model which is correctly specified at a disaggregated level is estimated with aggregate data, it is not likely that community research satisfies this test. Any community model in which a researcher is tempted to include county dummies will yield biased results when estimated with county data, for example. Furthermore, even if a model which passed the specification test were developed, the results and their policy implications could not really be trusted. We have argued that, in the example of community research, inadvertent theory and policy recommendations have been developed because of this latter consequence.

REFERENCES

Alexander, Karl L. and Larry J. Griffin. "School District Effects on Academic Achievement: A Reconsideration." *American Sociological Review* 41(1976):144-152.

Ball, D. S. and J. W. Wilson. "Community Health Facilities and Services: The Man Power Dimensions." *American Journal of Agricultural Economics* 50(1968):1208-22.

Bidwell, Charles E. and John D. Kasarda. "Reply to Hannan, Freeman and Meyer, and Alexander and Griffin." *American Sociological Review* 41(1976):152-160.

Bonjean, Charles M., Harley Browning and Lewis Carter. "Toward Comparative Community Research: A Factor Analysis of United States Counties." *Sociological Quarterly* 10 (1969):157-176.

Breisch, W. F. "Impact of Medical School Characteristics on Location of Physical Practices." *Journal of Medical Education* 45(1970):1068-1070.

Coleman, Sinclair. *Physician Distribution and Rural Access to Medical Services:* R-1887-HEW. Santa Monica, California: The Rand Corporation, 1976.

Feige, Edgar L. and Harold W. Watts. "An Investigation of the Consequences of Partial Aggregation of Micro-Economic Data." *Econometrica* 40(1972):343-60.

Firebaugh, Glenn. "A Rule for Inferring Individual-Level Relationships From Aggregate Data." *American Sociological Review* 43(1978):557-572.

Hannan, Michael T., John H. Freeman and John W. Meyer. "Specification of Models for Organizational Effectiveness." *American Sociological Review* 41(1976):136-143.

Kaufman, Harold F. "Toward an Interactional Conception of Community." *Social Forces* 38(1959):8-17.

Kaufman, Harold F. and Kenneth P. Wilkinson. *Community Structure and Leadership: An Interactional Perspective in the Study of Community.* State College: The Mississippi State University, Social Science Research Center, Bulletin 13, 1967.

Robinson, W. S. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review* 15(1950):351-357.

Scheffler, R. M. "The Relationship Between Medical Education and the Statewide Per Capita Distribution of Physicians." *Journal of Medical Education* 46(1971):995-998.

Wilkinson, Kenneth P. "The Community as a Social Field." *Social Forces* 48(1970):311-322.

Wilkinson, Kenneth P. "A Field Theory Perspective for Community Development Research." *Rural Sociology* 37(1972):43-52.