7-1-1980

Definitions of community: an illustration of aggregation bias, Station Bulletin, no.516

Luloff, A. E.

Greenwood, P. H.

New Hampshire Agricultural Experiment Station

'ION BULLETIN 516 July, 1980

Definitions of Community: An Illustration of Aggregation Bias

by

A. E. Luloff and P. H. Greenwood

NEW HAMPSHIRE
AGRICULTURAL EXPERIMENT STATION
UNIVERSITY OF NEW HAMPSHIRE
DURHAM, NEW HAMPSHIRE

ACKNOWLEDGEMENTS

This publication is a result of the research program of the Institute of Natural and Environmental Resources. The Institute is a multi-disciplinary group of scientists involved in a coordinated program of research, teaching and extension. The research effort encompasses investigations of: problems affecting the quality of the environment, economics of agriculture, forest and wildlife resources, the efficient use and conservation of water and soil, and regional and community planning and development.

The authors wish to acknowledge the assistance of Glenn Israel, Tom Frisbee and Tom Ilvento in data collection and organization. This paper represents a contribution to Regional Research Project NE-129 and to Project H-266 of the New Hampshire Agricultural Experiment Station.

ABSTRACT

Although continued attention has been given to the general study of "community," we still lack a consensus, operational definition. This absence impedes development of a unified sociology of community. Because authors have used different area conceptualizations, knowledge is, at best, case specific. Our examples demonstrate how similar conceptual models estimated with different community definitions generate divergent levels of statistical and substantive significance. Such findings underlie the need for social scientists to pay more careful attention to their areal definitions when study "community."

KEY WORDS: Sociology, Aggregation Bias, Community, Tax Characteristics

TABLE OF CONTENTS

Definitions of Community: An Illustration of Aggregation Bias

by

A. E. Luloff and P. H. Greenwood*

INTRODUCTION

Aggregation belongs to the class of empirical problems which may be assumed away. While assumptions tend to be painless, they are normally not costless. The presence of aggregation problems may bias one's results and distort one's hypothesis tests regardless of whether or not they are assumed away. Potential problems with aggregation arise for many reasons. One important cause is a research interest that defies consensus definition. For example, there are researchers interested in the behavior of community and while these researchers are prepared to offer a definition of a community, they are hard pressed to find an operational analog. Similarly, marine economists may consider models of a fishery when a fishery is no less abstract a concept than a community. If a county is regarded as a collection of communities the problem of defining a community is avoided provided that aggregation problems are assumed away. Those familiar with the quasi-community literature will not be surprised at the number of data collection units which have been used as either community surrogates or aggregates. In the absence of any aggregation problems, we should be indifferent as to the unit over which data is collected. Simply assuming a problem away is not the most satisfying method for dealing with a problem that poses real hazard of distoring one's perception.

The Problem of Aggregation

Applied research on the nature of community typically involves the empirical determination of the relations among a set of variables, and the comparison of these results with theoretically derived hypotheses. It should be clear that, if the comparisons are to be meaningful, the empirical and theoretical results must be compara-

*Assistant Professor of Community Development and Assistant Professor of Resource Economics, respectively, Institute of Natural and Environmental Resources, University of New Hampshire, New Hampshire Agricultural Experiment Station, Durham, NH.

ble. This comparability is sometimes lost when the empirical results are derived from aggregated (or disaggregated) observations on the variables. For example, a theory may imply a linear relationship between a set of independent variables, one of which may be stochastic, and a dependent variable; the relationship may be estimated by minimizing the sum of squared residuals. If aggregated observations are used in the estimation, the regression coefficients may suffer from bias, that is, their expectations may differ from their theoretical counterparts. Therefore, when county data is used to test a community hypothesis, aggregation bias is a potential problem. This problem evaporates if the county and the community are coincident, and if this is not a viable assumption then the problem would evaporate if the county averages are uncorrelated with the stochastic elements of the community observations (cf. Firebaugh, 1978).

Aggregation bias has long been recognized as a problem. More recently it has been shown that aggregation interferes with the application of the t test of these coefficients, and furthermore may play havoc with the measure known as R^2 (Greenwood and Luloff, 1979). These impacts do not necessarily require the preconditions for bias. Therefore, if the comparability between theoretical and empirical findings is assumed incorrectly, unsupportable hypotheses may find support, and the confidence in the predictive power of the theory may be falsely bolstered. That these are among the consequences of aggregation may be shown theoretically, but the practical significance can, perhaps, be best indicated by example.

Two approaches to demonstrating the confounding impact of aggregation suggest themselves. The first is an arbitrary approach in which a set of observations is transformed by arbitrary rules into sets of aggregated observations. Each set of transformed observations could be used to estimate the regression coefficients, the t values, and the value of R^2. The major advantage with this approach is that it would be inexpensive to implement; the drawback is that the transformation rules are arbitrary. In practice, transformation rules are not arbitrary. County data, for example, aggregates minor civil divisions which have something in common; they are contained within the same county. A second approach would be to look at real data and transform it into aggregated observations using accepted aggregate concepts. This is relatively more expensive since a large data base is needed. Moreover, the number of aggregations is restricted. Another difficulty is that there is no clear benchmark with this approach. Arbitrary data may be determined with known coefficients and stochastic parameters. Since the reason for the examples is to demonstrate the confounding

impacts of aggregation, the lack of a clear benchmark is not a major drawback. It is enough to show the variability in the results without indicating which set of results is somehow best.

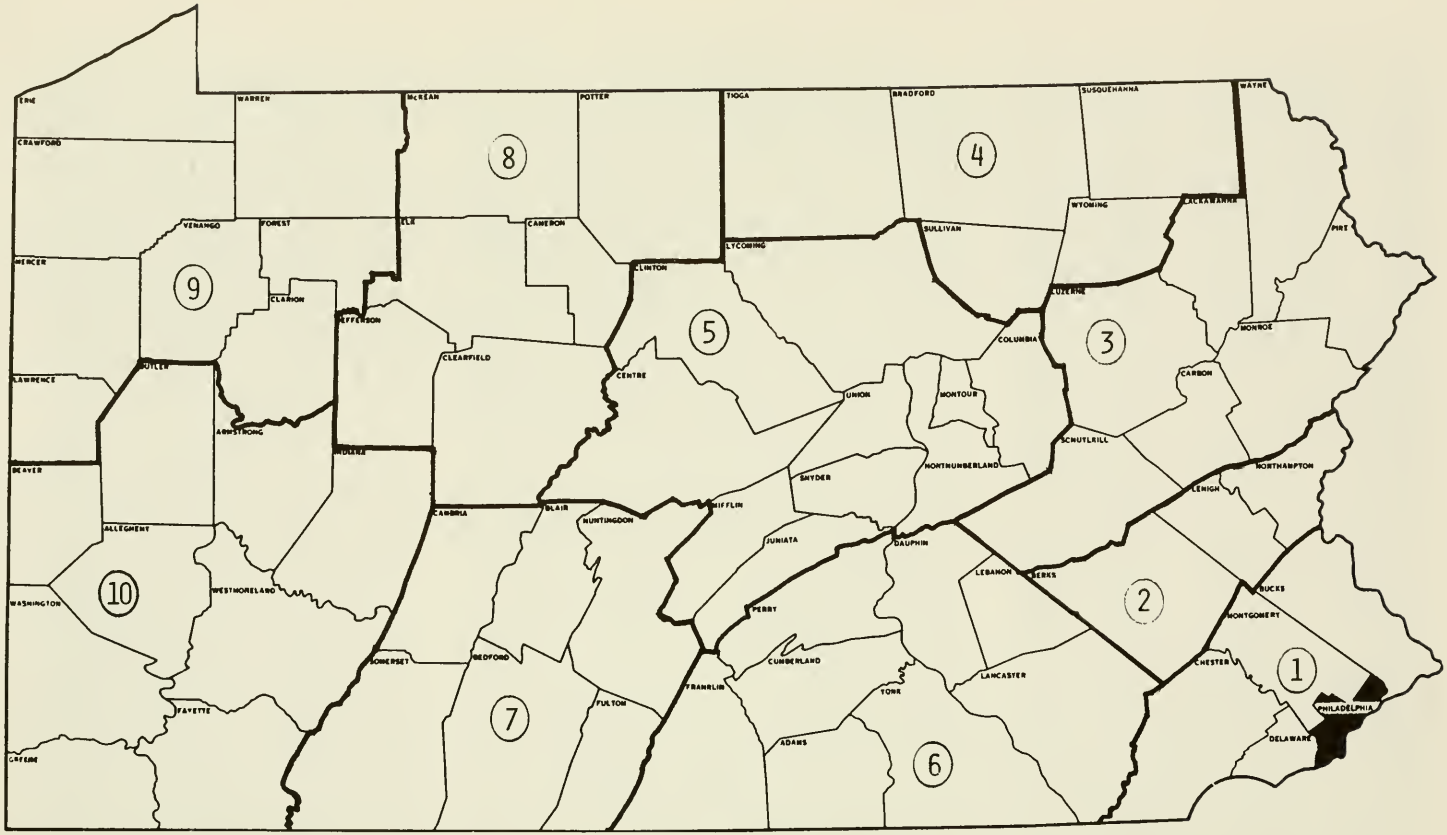
Since a reasonably large set of data was available to us at low cost, the second approach was chosen. We estimated three models at three levels of aggregation; the same observations were used in each case. These examples will demonstrate that the confounding consequences of aggregation are not idle prospects that can be ignored with impunity.

Pennsylvania provides the setting for these examples. Observations on more than 2,000 minor civil divisions (MCD's) were collected. These observations were collapsed into 66 county observations, and these in turn were collapsed into observations on ten regions (see figure 1).[1] We generated regression results at each level of aggregation.

The first model attempts to identify characteristic patterns of local, county, and regional tax behavior. In Pennsylvania, localities are entitled to impose a number of taxes, other than real estate and occupation taxes, on its residents (the state imposes a sales and income tax on all residents). This right derives from the Local Tax Enabling Act (Act 511) of 1965 (commonly referred to as the "Tax Anything Law"). Included within the categories of taxes allowed through this legislation is a per capita head tax. To account for the level of 1974 per capita tax revenues we selected two variables: (1) the level of these revenues in 1970; and (2) the change in earned income derived taxes between 1970 and 1974. The first variable provides a historical benchmark, and the second provides a measure of the shift in dollars generated through the exercise of a 511 tax.

Table 1 presents three sets of results for this simple model. Each row contains the estimated constant (α), the estimated coefficient on the 1970 level of per capita tax (B_1), the estimated coefficient on the

[1]In Pennsylvania there are 2,547 political subdivisions (excluding, counties, school districts, and authorities). Data were gathered for 2,463 municipalities. The remaining cases were eliminated from the analysis for several reasons. First, Pittsburgh and Philadelphia were eliminated on the basis of their uniqueness (by far the two largest metropolitan cities in the state). The elimination of Philadelphia also reduces the number of counties from 67 to 66 since Philadelphia is a county-city administrative unit. Second, many municipalities were eliminated because they were involved in political mergers with other municipalities or because their census identification numbers did not match with other sources of data. The remaining cases were deleted because census data were undisclosed for these communities. The availability of a data set which includes 96.7% of all municipalities, 66 counties, and affords us the opportunity to use the 10 uniform regions so designated by the Pennsylvania Department of Community Affairs prompted us to adopt the second approach described above. Further, because of the makeup and distribution of its population, Pennsylvania is often used to generalize to the country as a whole (cf. Zelinski, *et al.*, 1974).



■ = Deleted from Analysis

Figure 1. Pennsylvania's Counties and Regions

change in earned income tax (B_2), the value of R^2 adjusted for degrees of freedom, and the sample size (n). Values of the t statistic appear in parentheses. Each column contains a regression result at the MCD level, the county level, and the regional level respectively. This format will be followed in later tables. The same basic data is used for each of the three models; minor civil division data is aggregated into its respective county units and the regional data is derived by aggregating the appropriate county units. The observation of the county level is the mean ($\overline{X}$) of the MCD level data; likewise the regional mean is an average of its constituent county means.

Table 1. Summary of Regressions at Three Levels of Aggregation for Equation 1.

Level of Agregation	α	B_1	B_2	R^2Adj.	n
Minor Civil Division	1990	.72* (51)[1]	-.021* (6.8)	.55	2463
County	599	.93* (14)	-.0025 (.17)	.86	66
Region	583	.96* (17)	-1.43 (.35)	.97	10

*Significant at .001.

[1]Numbers in parentheses are simple t statistics.

The coefficient B_1 is positive, less than one, and significant in each of the three regressions. The coefficient B_2 behaves a little more peculiarly. In all cases it is negative; in the first two it is fairly small and in the third it is quite large. At the regional level our best estimate is that for every dollar increase in earned income taxes, per capita tax collections fall by $1.43. However, only at the MCD level is B_2 found to be significantly different than zero. Looking at either the county or regional results (only), it is clear that the hypothesis, $B_2 = 0$, could not be rejected. In theory this is the same B_2 that exists at the MCD level or at the "community" level. Yet at the MCD level it is clear that we would reject the hypothesis that $B_2 = 0$. The issue is not whether one conclusion or the other is more correct; our point is simply that our conclusion depends in part on the level of aggregation that we select. Even when we adjust for the loss of degrees of

freedom, there is seen a pattern of increasing R^2 with the level of aggregation. This is an expected pattern but not universal. With the MCD equation both variables are significant and R^2 indicates that there are a number of influences that have not been controlled. At the regional level, R^2 is so high that there do not appear to be many influences left to control for, and, indeed, changes in earned income taxes account for very little of the variance of the dependent variable. It is debatable whether the regional equation explains more than the MCD equation or simply hides more.

The second model investigated represents a different but related attempt at accounting for the observed level of 1974 per capita tax revenue. Two variables were again selected: (1) the change in earned income taxes (defined and utilized as in Model 1), and (2) the change in total Act 511 tax dollars collected. The results are presented in Table 2. Each row in this table includes the constant (α), the estimated coefficient on the change in earned income taxes (B_1), the estimated coefficient on the change in Act 511 collections (B_2), and the summary measures as before.

Table 2. Summary of Regression at Three Levels of Aggregation for Equation 2.

Level of Aggregation	α	B_1	B_2	R^2Adj.	n
Minor Civil Division	749	-.142* (24.8)[1]	.085* (18.8)	.209	2463
County	458	-.037 (1.26)	.017 (.83)	.007	66
Region	193	.096 (1.73)	-.066 (1.96)	.218	10

*Significant at .001.

[1]Numbers in parentheses are simple t statistics.

At the MCD level, changes in earned income taxes have a negative impact, while changes in Act 511 collections have a positive impact. Both of these are significant. One possible interpretation is that, as income grows, localities become less reliant absolutely on per capita taxation because other collections are positively related to income. Moreover, areas that have changed the level of Act 511

collections are likely to adjust their per capita collections in the same direction. The value of R^2 for the minor civil division regression is low (.209).

Not unexpectedly, major differences appear at the remaining levels of aggregation. At the county level, neither coefficient is significant, although their signs are the same as they were at the MCD level. Moreover, the county equation has virtually no predictive power. At the regional level, the equation has changed considerably. The signs on the coefficients have switched (from the directions of both the minor civil division and county level equations respectively). Further, R^2 has rebounded to its earlier level. However, unlike the first equation, neither coefficient is significant at the .001 level, although the t values are not small.

Obviously these are perplexing results. Despite the condition that the same data are used in all three cases, the empirical results are not generalizable. There is no apparent relation at the county level, and contrary relations exist at the MCD and regional levels.

The final model investigated attempts to account for the changes in total tax revenues collected during the period 1970-1974. The results are presented in Table 3. This change is viewed as a function of the changes in Act 511 taxes (B_1) and the changes in non-Act 511 taxes (B_2). The latter taxes are primarily generated

Table 3. Summary of Rgressions at Three Levels of Aggregation for Equation 3.

Level of Aggregation	α	B_1	B_2	R^2Adj.	n
Minor Civil Division	5009	1.38* (66)	-1.27* (5.4)	.64	2463
County	-4172	1.75* (24.3)	-.025 (0)	.901	66
Region	-7068	1.96* (10.5)	9.06 (.897)	.941	10

*Significant at .001.

[1]Numbers in parentheses are simple t statistics.

through real estate and occupation taxes. Real estate remains the chief local tax source for this state.[2]

The coefficient B_1 is positive, greater than one, and significant in all three regressions. However, there is considerable variation in coefficient B_2. At the MCD level the coefficient on the non-511 taxes is significant and less than negative one. At the county level, the coefficient is also negative, but it is not significant and it approaches zero. At the regional level, the coefficient reverses its sign, is much greater than one, and remains insignificant. Again, conclusions which are supported at one level of aggregation are clearly unsupportable at other levels.

This paper has provided an empirical demonstration of some of the problems inherent to aggregation which have been discussed elsewhere. We have observed cases where R^2 falls with increased aggregation, R^2 increases with increased aggregation, and R^2 remains relatively constant with increased aggregation. More significantly, the coefficients have switched signs and magnitudes, in some cases they have lost statistical significance, and in one case we have seen a sign switch direction while retaining significance (although at a lower level) as aggregation increased.

One of the questions motivating this exercise is how does one assimilate "community" research conducted at various levels of aggregation. Findings of significance at one level need not generalize to the "community." Results with high predictive power may obscure "community" realities, and the reverse may also be true. Situations in which no constructive results emerge also need not generalize to the "community."

As pointed out by Blalock (1979) in his presidential address to the American Sociological Association, this problem is endemic to social science research. It is of particular importance to the community researcher, however, because of the lack of consensus surrounding the definition of community. In instances where aggregation of data has occurred, due to decision criteria established by others, the researcher needs to be aware of the limited generalizability of his/her results.

---

[2]While we use both 511 and non-511 tax generated revenue in our models, we do not mean to imply that these are the only categories of tax sources. Indeed, non-tax revenues accounted for nearly 41% of all revenue generated in 1970. The source of this revenue includes dollars generated through public service enterprises, water supply and sewer charges, state and federal grants, licenses, permits, and fines.

REFERENCES

Blalock, H. M. 1979. "The Presidential Address: Measurement and Conceptualization Problems: The Major Obstacle to Integrating Theory and Research." American Sociological Review 44 (December):881-894.

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

Greenwood, Peter H. and A. E. Luloff. 1979. "Inadvertent Social Theory: Aggregation and Its Effect on Community Research." Journal of the Northeastern Agricultural Economics Council VIII (April):44-47.

Zelinsky, Wilbur, *et al.* 1974. Population Change and Redistribution in Nonmetropolitan Pennsylvania, 1940-1970. Pennsylvania State University: Population Issues Research Office.