GEOGRAPHICAL ANALYSIS, Vol. VIII (October 1976). Submitted 1/75. Accepted 4/75.

W. A. V. Clark and Karen L. Avery

The Effects of Data Aggregation in Statistical Analysis

Abstract

The problem of ecological correlation is now widely recognized but detailed analyses of the effects of aggregation on correlation and regression coefficients are rare. A short review of the aggregation problem is followed by an analysis of the specific effect of proximity aggregation on the slope coefficient of a bivariate linear model using data drawn from the Los Angeles Metropolitan region. The evidence suggests that changes in the slope coefficient are best related to the manner in which the covariation between the independent and dependent variables changes with increased aggregation.

The aggregation problem has been prominent in the analysis of data in almost all the social sciences and some physical sciences. In its most general form the aggregation problem can be defined as the information loss which occurs in the substitution of aggregate, or macrolevel, data for individual, or microlevel, data. A recent review of urban models discusses the impact of aggregation on the predictions from those models [5], and another study notes that with increasing level of aggregation more and more spatial frequencies may be lost and so obscure important processes in the system [7]. There is also loss of information when quadrats are aggregated prior to fitting probability distributions [4, p. 61]. A general review of the nature of the aggregation problem and the ways in which social scientists have evaluated its effect on their research results is contained in [8]. The present paper focuses on a

W. A. V. Clark is professor of geography and Karen L. Avery is a graduate student in geography, University of California, Los Angeles.

specific element of the aggregation problem, the occurrence of aggregation bias in correlation and regression analysis. In particular, this type of bias is manifested as the inflation of macrolevel coefficient estimates above the corresponding values of these coefficients estimated from microlevel data.

It has long been known that the use of aggregate data may yield correlation coefficients exhibiting considerable bias above their values at the individual level [*10, 21*]; and Blalock [*2*] has shown that the regression coefficients may be biased also. It is well established that it is incorrect to assume that relationships existing at one level of analysis will necessarily demonstrate the same strength at another level. The estimates derived from aggregate data are valid only for the particular system of observational units employed. The consequences of using potentially biased estimates of the correlation and regression coefficients as substitutes for the "true" microlevel estimates are most serious in terms of the causal inferences to be drawn from statistical analyses.

A variety of technical and theoretical explanations of the so-called aggregation problem in correlation and regression (and some solutions) have been offered during the past two decades, but the specific aggregation conditions under which bias can be expected to occur require greater understanding, particularly for situations in which the predominant criterion for aggregation is the spatial proximity of observational units. Much statistical analysis in the social sciences involves the use of observational units which are aggregates of smaller units. In some instances of practical research, recorded data are available only in aggregate form and in others data collection is constrained by time or cost limitations, but in any case social scientists in general and geographers in particular continue to work with aggregate data as a surrogate for individual data.

Probably the most serious disadvantage of using aggregate data is the inherent difficulty of making valid multilevel inferences based on a single level of analysis [*1*]. Alker has identified three types of erroneous inferences that may appear should a researcher attempt to generalize from one level of investigation to another. The individualistic fallacy is the attempt to impute macrolevel (aggregate) relationships from microlevel (individual) relationships. It is the classic aggregation problem first examined by economists, and according to Hannan [*15*, p. 5] it "concerns attempts to group observations on 'behavioral units' so as to investigate economic relationships holding for sectors or total economies." Cross-level fallacies can occur when one makes inferences from one subpopulation to another at the same level of analysis. The ecological fallacy, so named from the work of Robinson [*18*], is the opposite of the individualistic fallacy and involves making inferences from higher to lower levels of analysis. Robinson demonstrated that there was not necessarily a correspondence between individual and ecological correlations, and that generally the latter would be larger than the former. Although the ecological fallacy has been widely discussed and publicized, it is still a common error in studies involving causal inference.

In studies in which one is interested primarily in the causal analysis of microlevel rather than macrolevel relationships, and yet is obliged to work with aggregate data, one should attempt to employ a system of data grouping that produces as little loss of information on the individuals as possible. Thus the ideal aggregation procedure would yield groups which are homogeneous with respect to all of the variables in the model. However, determining the optimal grouping procedure is not straightforward. Systematic grouping is likely to yield biased, albeit relatively efficient, estimates, and random grouping tends to produce unbiased, but inefficient, estimates [*15, 6*]. The use of data recorded as summary measurements for individuals within areal units can be viewed as a special case of systematic grouping, in which the main criterion for grouping is geographical proximity. Such areal units may in fact be relatively homogeneous internally due to the spatial autocorrelation of many socioeconomic phenomena. However, spatial proximity as a grouping criterion may or may not yield homogeneous units, depending on the size of the units and the degree to which the variables of interest exhibit spatial autocorrelation.

Robinson's assertion that ecological correlations can never be used validly as substitutes for individual correlations stimulated many attempts to develop techniques whereby individual parameters could be derived from ecological parameters. Duncan and Davis [*9*] proposed a method of determining upper and lower bounds for the true individual correlation from aggregate data, and Goodman [*11, 12*] postulated some special circumstances in which the regression between ecological variables could be used to make inferences about the behavior of individuals. However, none of the proposed techniques presented a comprehensive solution. Robinson's paper provided a generally satisfactory, acceptable explanation of the mathematical relationship between microlevel and macrolevel correlations, but it became apparent that solutions to the problem were being attempted without a sound theoretical understanding of exactly how aggregation affects variation in the independent and dependent variables.

Blalock has attempted to specify more clearly the varying nature of the aggregation procedure, particularly in the context of linear causal analysis. He has examined four ways in which grouping procedures may affect bivariate causal relationships: (1) random grouping, (2) grouping to maximize variation in the independent variable, (3) grouping to maximize variation in the dependent variable, and (4) grouping on the basis of spatial proximity of individuals [*2*, pp. 104–12]. The last procedure, aggregation of observations on the basis of spatial proximity, is the case most likely to be of particular interest to geographers.

Economists have also examined the specific mathematical relationships between various grouping procedures and their effects on regression coefficients [*19, 20*]. Prais and Aitchison [*17*] show that grouping by the independent variable will yield unbiased regression parameters, though these may not be efficient estimates. Cramer [*6*] shows that as the size of the aggregate unit increases, the efficiency of the macrolevel

estimate of the regression coefficient decreases.

To illustrate the problems particularly related to the use of census tract data, and to demonstrate the specific effect of proximity aggregation on the slope coefficient of a bivariate linear model, the remainder of the paper reports on an empirical investigation of data from the Los Angeles region.

Empirical Analyses of Aggregation Effects

The hypothesized model is of the general linear form

$$Y_i = a + bX_i + e, \tag{1}$$

where Y is a measure of family income and X is a measure of the level of education of the head of household. The first part of the study compares individual and census tract groupings of data, and the second focuses on groupings of individual households. The individual household data were obtained from the Los Angeles Metropolitan Area Survey (LAMAS) conducted in 1972 by the Survey Research Center at UCLA. A cluster sample of 1,024 households yielded 952 usable interviews. Additional data were collected for 1,556 of the 1,596 census tracts in Los Angeles County for 1970. The LAMAS income data was originally reported in fifteen classes, and for this study the midpoint of the class interval is used as the family income for each household. Level of education was recorded as the highest completed grade for the head of the household. For the census tracts, mean annual family income per tract and median number of school years completed by the head of household are used. While it would have been desirable to use mean values for both variables, the difference in measurement values is not crucial in the present study as we are interested primarily in the relative change in the desired coefficients. For the tract aggregations the group means of the aggregated tract values are used. In addition to the 952 individual units and 1,556 census tract units, two governmental groupings, the 134 Welfare Planning Council Study areas, and the 35 Regional Planning Commission Statistical Areas were used as aggregate units.

The estimates of the regression coefficient b, the product-moment correlation coefficient r, and the coefficient of determination r^2 are reported in Table 1. In that this study is not concerned with making inferences to a larger population, the assumptions of the regression model are not of paramount importance. However, there is some chance that systematic heteroscedasticity biases might influence the magnitude of the correlation coefficients independently from the effect of the aggregation procedure. To examine this proposition, scatter diagrams and plots of residuals against the dependent variable at all levels of analysis were investigated. In the case of the unaggregated tracts, and the one of the groupings of individuals from the LAMAS data there was evidence

TABLE 1

CORRELATION AND SLOPE COEFFICIENTS DERIVED FROM LAMAS AND CENSUS TRACT DATA

Data Set	Method of Data Generalization	r	r^2	b
952 units: LAMAS	Not applicable	0.4028	0.1623	857.60
1,556 units: Census tracts	Tract mean	0.6434	0.4140	2413.64
	Tract mean (log income)	0.6707	0.4498	0.0892
134 units: Welfare Planning Council	Group mean	0.7606	0.5785	2808.21
	Group mean (log income)	0.8285	0.6864	0.1051
35 units: Regional Planning Commission	Group mean	0.8503	0.7230	3103.62
	Group mean (log income)	0.8811	0.7763	0.1088

of increasing variance with increasing variable magnitude. It is unlikely that these isolated results affect the overall conclusions to be derived from the several levels of analysis employed in this study. However, the income and education data for tracts, and the 134 and 35 groups of tracts, are best fitted by the relationship $\log Y_i = a + bX_i + e$. The results for this model are reported in Table 1, but again they do not affect the conclusions to be drawn from the study.

Inspection of the correlation coefficients reveals that in general the variations in r and r^2 caused by aggregation of the data conform quite well to the expected results predicted by Hannan and Blalock [*2, 3, 15*]. Aggregation of observational units on the basis of proximity leads to substantially biased correlation coefficients, with an increase in r as the level of grouping increases. Whereas r and r^2 are fairly low at the microlevel (0.4028 and 0.1623 respectively), the macrocoefficients obtained at the tract level are substantially higher ($r = 0.6434$ and $r^2 = 0.4140$). Data derived from the proximity aggregations of tracts (i.e., the 134 Welfare Planning Council Study Areas and the 35 Regional Planning Commission Statistical Areas) yielded macrocoefficients which are even higher than those obtained at the tract level. The system of tract aggregations having fewer groups (35) tended to produce correlation coefficients higher than those derived from the system consisting of 134 groups.

The variation in the regression coefficients obtained from the different levels of data aggregation is not entirely consistent with the previous findings of Blalock [*2*, p. 103]. He reported that the aggregation of data on the basis of proximity did not appear to effect significant differences between the micro- and macrolevel regression coefficients b obtained from the particular set of data used. As has been stated previously, Blalock attributes any change, or lack of change, in the regression coefficients derived from proximity grouping to the degree to which the contiguity criterion for grouping systematically affects variation in the independent or dependent variable. If grouping maximizes variation in the independent variable, b should remain the same; but if variation in the dependent variable is maximized, the slope should increase (see Hannan [*13*, p. 47]). Further, Hannan feels that it is an empirical question "whether or not proximity . . . grouping will have

more effect on variation in X or Y." At the individual level in the current study, the estimate of the slope coefficient for the LAMAS data is 857.60 (see Table 1). For the ungrouped tract data, b is 2413.64. At higher levels of proximity aggregation (i.e., the Welfare Planning Council and Regional Planning Commission tract groupings), b tends to increase, in contrast to Blalock's findings.

However, certain difficulties inherent in the data place some restrictions on the extent to which further comparative analyses of this data can be made. Perhaps the most obvious problem involves the comparison of the individual level data with the tract level and aggregated tract data. Since the census tracts are not true aggregates of the 952 individuals of the LAMAS data, it may not be valid to make more than cursory comments on the relationships between coefficients derived at the microlevel and those at the various macrolevels. The problems of data generalization (specifically the use of group means in the aggregated data) which may have biased the results and the difficulty of making any reliable statements about how aggregation produces changes in the variation of the independent and dependent variables and their covariation when unequal-sized groups are used are more serious. To overcome these difficulties the subsequent analysis uses only the LAMAS data, and imposes certain restrictions on the aggregation procedure.

Proximity Aggregation and Biased Regression Coefficients

As already noted, Blalock [*2*, p. 111] found in one case study that proximity aggregation did not produce bias in the slope coefficient at the macrolevel. Hannan [*13*, p. 47], extrapolating from Blalock's work, has implied that for any situation involving proximity grouping, the existence or lack of bias in b at the macrolevel is dependent upon whether the aggregation criterion has a greater effect on the variation in the independent *or* dependent variable of the bivariate model. It is expected that aggregation that maximizes variation in Y will lead to an increase in b, whereas maximization of the variation in X will yield an unbiased slope coefficient at the macrolevel. Hannan contends that proximity aggregation may lead to either of these two conditions, contingent on the nature of the data employed in any particular study. Thus the crux of the issue, as expounded by Hannan and Blalock, seems to lie in whether proximity aggregation maximizes variation in X relative to that in Y, or vice versa.

To examine this hypothesis, data derived from the Los Angeles Metropolitan Area Survey are used. In addition to the microlevel of 952 households, four levels of aggregated observational units have been devised. The dominant criterion for aggregation is spatial proximity, and each aggregate unit has been made as spatially compact as possible. Within any one level, each group consists of an equal number of individual households. The first aggregate set of data is composed of 136 groups

TABLE 2

DATA DERIVED FROM AGGREGATION OF LAMAS HOUSEHOLDS

	DATA SET				
	952 Units	136 Units	68 Units	34 Units	17 Units
$\bar{X}_k$	12.44223	12.44223	12.44302	12.44219	12.44220
$\bar{Y}_k$	10688.54688	10688.34375	10688.50781	10688.50391	10688.47266
BSS_X	10,055.6343	3513.9048	2364.5020	1529.2956	1178.9512
BSS_Y	45,581,569,417.5617	15,376,615,458.3993	10,267,180,739.7292	6,794,268,818.8956	4,859,257,965.4864
BCP_{XY}	8,623,705.0890	4,236,301.7550	3,297,064.3720	2,325,527.8200	1,678,740.224
r	0.4028	0.5763	0.6692	0.7214	0.7014
r^2	0.1623	0.3321	0.4478	0.5205	0.4919
b	857.5981	1205.5886	1394.4004	1520.6567	1423.9172

of 7 individuals each. The other groups are: 68 groups of 14 individuals, 34 groups of 28 individuals, and 17 groups of 56 individuals.

As before, a bivariate linear model, where Y is total annual family income and X is the highest grade in school completed by the head of household, is tested. Least squares and product moment procedures were used to estimate the coefficients of regression and correlation. Within each set of aggregated data, the calculations of the coefficients have been carried out using the mean of the group means, rather than the grand mean of the individuals. Since equal-sized groups are employed, the group means in each case are virtually identical to the grand mean (Table 2). The total sums of squares of X and Y (denoted as BSS_X and BSS_Y) and the total sum of cross products of X and Y (denoted as BCP_{XY}) at the microlevel, as well as the between-groups sums of squares and cross products at the macrolevels, are presented in Table 2.

As expected, the correlation between income and education tends to increase markedly with the level of aggregation. Whereas only a weak positive relationship is exhibited at the microlevel ($r^2 = 0.1623$), X explains approximately 50 percent of the variation in Y at the two highest levels. Similarly, the slope coefficient b also tends to increase. An apparent anomaly occurs at the fifth level, as both b and r show decreases below their corresponding values at the fourth level.

In terms of the slope coefficient, we would expect from the previous studies of Hannan and Blalock that b would show no bias at the macrolevels, since, as in the case of random grouping, aggregation has affected the variation in both variables to nearly the same relative degree. The fact that aggregation slightly maximizes the between-groups variation in X relative to that in Y serves only to reinforce the perplexity of this departure from the norm, since Blalock has hypothesized that maximization of the variation in X relative to the variation in Y yields unbiased slope coefficients. However, both variables show a steady decrease in the between-groups sums of squares through all five levels, with no reversal of the trend between the fourth and fifth levels. Interestingly, it appears that both the variation in X and that in Y are decreasing at the same rate. Table 3 shows the ratio of the between-groups sum of squares to the total sum of squares for both X and Y (denoted as BSS_X/TSS_X and BSS_Y/TSS_Y) at each level. It is readily apparent that aggregation affects the variation in both variables to nearly the same

TABLE 3

VARIATION AND COVARIATION RATIOS DERIVED FROM AGGREGATION OF LAMAS HOUSEHOLDS

Data Set	BSS_X/TSS_X	BSS_Y/TSS_Y	BCP_{XY}/TCP_{XY}
952 units	1.0	1.0	1.0
136 units	0.34945	0.33734	0.49124
68 units	0.23514	0.22525	0.38233
34 units	0.15208	0.14906	0.26967
17 units	0.11724	0.10661	0.19467

relative degree at each level, the difference in the degree of reduction being less than 1.3 percentage points at all four macrolevels.

While scant explanation of the bias in b has been gained by examining the variation in X and Y, consider instead the covariation of X and Y. The total covariation of grouped data can be expressed as the sum of the within-group sum of cross products and the between-groups sum of cross products. In this study, the covariation of X and Y decreases as the level of aggregation increases. Note that at each macrolevel the ratio of the between-groups covariation to the total covariation (Table 3, denoted as BCP_{XY}/TCP_{XY}) is greater than the ratios of the between-groups variation to the total variation for either X or Y. Since the covariation in X and Y remains large relative to the variation in X at the various macrolevels, it follows that b should increase above its value at the microlevel. However, while the rates of change of variation in X and Y are initially higher than the rate of change of the covariation, this trend reverses itself at the fourth level. Between the fourth and fifth levels, the relative reduction in the covariation is about twice as great as the reductions in the variations of X and Y; between the first and second levels, the reductions in the variations of X and Y are greater than that in the covariation. Thus at the fifth level, b decreases below its value at the fourth level.

It would seem that the key to the conditions under which proximity aggregation produces bias in the correlation and regression coefficients lies in the understanding of how X and Y vary *together* and how aggregation directly affects that covariation, rather than in how X and Y vary separately. For this body of data, it does not seem unreasonable to assume that the socioeconomic characteristics of the population may be spatially autocorrelated. Low-level aggregation has apparently had the effect of grouping together high values of X and Y and, in other groups, low values of X and Y. At a lower level of grouping the groups are relatively heterogeneous with respect to one another, and the covariation remains large. When larger regions are formed, the effect of spatial autocorrelation decreases. The result is increasing internal heterogeneity within the groups, and a decreasing heterogeneity between the groups. This can be likened to random grouping, in which the covariation of X and Y is reduced in the same proportion as the variation in X or Y. The largest regions are approaching a random grouping situation, and, therefore, the covariation decreases at a relatively faster rate with increased consolidation of the data. It can be hypothesized that at even higher levels of aggregation, the reduction in the covariation would be in the same proportion as the reduction in the variations of X and Y and that b and r would approach their microlevel values.

Conclusion

The spatial aggregation of data has significant consequences in the correlation and regression analysis of areally distributed phenomena.

For the specific set of spatially aggregated data in this study, the value of the correlation coefficient r, derived from a bivariate linear model, tends to increase above its value at the microlevel, and there is also a concurrent increase in b, as the level of aggregation increases. However, these trends in the behavior of the coefficients do not remain constant through all levels of aggregation. The empirical evidence of this study does not support the Blalock hypothesis whereby the changes in the slope coefficient are explained by the reduction in the variation of the independent or dependent variable. An alternate hypothesis suggests that the deviations of the observed from the expected behavior of the coefficients are related directly to the manner in which the covariation between the independent and dependent variables changes with increased aggregation, and indirectly to the way in which spatial autocorrelation is exhibited among the micro- and macrolevel data. The implication for any statistical analysis of census tract information, including factorial ecological studies, is that the substantive conclusions should be treated with caution. Even when the census tract correlations are not used to infer individual relationships, the fact the coefficients are inflated by an unknown magnitude suggests that we recognize explicitly the possibility of bias in our analyses.

LITERATURE CITED

1. ALKER, R. JR. "A Typology of Ecological Fallacies." In *Quantitative Ecological Analysis in the Social Sciences*, edited by M. Dogan and S. Rokkan, pp. 69-86. Cambridge, Mass.: MIT Press, 1969.
2. BLALOCK, H. *Causal Inferences in Nonexperimental Research.* Chapel Hill: University of North Carolina Press, 1964.
3. ———. "Aggregation and Measurement Error." *Social Forces*, 50 (1971), 151-65.
4. CLIFF, A., and J. K. ORD. *Spatial Autocorrelation.* London: Pion, 1973.
5. COLENUTT, R. J. "Building Models of Urban Growth and Spatial Structure." *Progress in Geography*, 2 (1970), 109-52.
6. CRAMER, J. S. "Efficient Grouping, Regression and Correlation in Engel Curve Analysis." *Journal of the American Statistical Association*, 59 (1964), 233-50.
7. CURRY, L. "A Note on Spatial Association." *Professional Geographer*, 18 (1966), 97-99.
8. DOGAN M., and S. ROKKAN. *Quantitative Ecological Analysis in the Social Sciences.* Cambridge, Mass.: MIT Press, 1969.
9. DUNCAN, O. D., and B. DAVIS. "An Alternative to Ecological Correlation." *American Sociological Review*, 18 (1953), 665-66.
10. GEHLKE, C. E., and K. BIEHL. "Certain Effects of Grouping upon the Size of the Correlation Coefficient in Census Tract Material." *Journal of the American Statistical Association Supplement*, 29 (1934), 169-70.
11. GOODMAN, L. A. "Ecological Regressions and Behavior of Individuals." *American Sociological Review*, 18 (1953), 663-64.
12. ———. "Some Alternatives to Ecological Correlation." *American Journal of Sociology*, 64 (1959), 610-25.
13. HANNAN, M. T. *Aggregation and Disaggregation in Sociology.* Lexington, Mass.: D. C. Heath, 1971.
14. ———. "Problems of Aggregation." In *Causal Models in the Social Sciences*, edited by H. Blalock. Chicago: Aldine, Atherton, Inc., 1971.
15. ———. *Approaches to the Aggregation Problem.* Technical Report No. 46, Laboratory for Social Research, Stanford University, 1972.

16. MENZEL, H. "Comment on Robinson's 'Ecological Correlations and the Behavior of Individuals'." In *Studies in Human Ecology,* edited by George A. Theodorson, pp. 121-22. Evanston, Ill.: Row, Peterson, 1961.
17. PRAIS, S. J., and J. AITCHISON. "The Grouping of Observations in Regression Analysis." *Review of the International Statistical Institute,* 22 (1954), 1-22.
18. ROBINSON, W. S. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review,* 15 (1950), 351-57.
19. THEIL, H. *Linear Aggregation in Economic Relations.* Amsterdam: North Holland, 1954.
20. ———. "Alternative Approaches to the Aggregation Problem." In *Logic, Methodology and the Philosophy of Science,* edited by E. Nagel, P. Suppes, and A. Tarski, pp. 507-27. Stanford: Stanford University Press, 1960.
21. YULE, A. U., and M. G. KENDALL. *An Introduction to the Theory of Statistics.* New York: Hafner Publishing Company, 1950.