APPROACHES TO THE AGGREGATION PROBLEM

by

Michael T. Hannan
Stanford University

Technical Report No. 46 1972

ABSTRACT

This paper attempts to further development of a general "theory of aggregation" which will integrate the results developed for specialized social science research applications. We first formulate a general model within which aggregation bias is defined. Given this formulation three well developed perspectives on the methodological problems of aggregation are compared -- a "classical" grouping approach, a causal models approach and a specification error approach. All three perspectives are reasonably useful for the simplest cases. However, the causal models and specification error approaches are preferable as general formulations since they deal more adequately with realistic complications. No existing approach handles aggregation in multivariate models in a completely satisfactory manner. A number of suggestions are made for extending existing formulations to remedy this situation.

APPROACHES TO THE AGGREGATION PROBLEM*

The term aggregation is used loosely in the social sciences to refer to a broad class of rather diverse issues. It refers to conceptual and theoretical issues involved in attempts at composition or shifts in levels of analysis. It is likewise used to refer to attempts at index construction or data reduction where sets of indicators or variables are combined. Finally, the term is used to refer to analysis issues involving shifts in levels of data aggregation. While the theoretical and methodological issues have implications for each other, it is unlikely that we can develop a single abstract calculus for analyzing both types of issues.[1] There do exist, however, abstract formulations which develop of the partial similarity of the aggregation of variables and the aggregation of observations. Yet, at the same time these formulations also clarify the important differences in the methodological issues arising in the two cases. And, it is safe to conclude for the present that each of these cases should be analyzed separately.

This paper deals only with approaches to the last named type of aggregation issue. In nontechnical terms we will say that aggregation problems arise whenever an analyst makes inferences from a model estimated at one level of data aggregation to properties of an analagous model at a different level of aggregation.

Our usage of the term aggregation problem is still broad enough to admit into consideration a considerable number of methodological and analysis issues. And, in fact social scientists have encountered variants of the more narrowly defined aggregation problem in a wide variety of research situations. The concrete features of the applications are so different that there has been a marked tendency for specialized and discrete methodological literatures to develop around each version of the problem. While this trend has resulted in

a rich variety of special results, it seems not to have improved our understandings of the general features of the problem. There have been a number of attempts at more encompassing formulations (e.g. Theil, 1954; Blalock, 1964; Hannan, 1971). But, there are important areas in which the general formulations and the specialized results are not well articulated. This paper attempts to clarify a number of such issues. In this attempt we survey a number of fairly well developed approaches to the aggregation problem. The dominant concern is with extending and improving the general formulations and with drawing implications for research practice.

Section I outlines five practical research situations in which aggregation issues typically arise. Then Section II abstracts a common framework from the concretely different situations and develops a consistency formulation within which to assess the consequences of aggregation. The development of the formal apparatus continues in Section III where aggregation bias is defined. Section IV analyzes three different approaches to the problem of aggregation for the bivariate case. The issues in this case have been well studied by now and our knowledge is fairly complete. But, the extension of these approaches to the multivariate case is enormously complicated. Since the practical importance of this methodological work depends heavily on the ability to generalize to multivariate models, we devote considerable attention to this problem. Section V develops a formal model of specification bias and treats aggregation bias as a special case. The utilization of the specification error apparatus is not uncomplicated as we shall see and in the closing section we comment on the practical use of "aggregation theory" and outline a number of important unsolved problems.

I CLASSES OF PRACTICAL AGGREGATION PROBLEMS

Before moving on to abstract formulations of the aggregation problem, we will briefly describe five quite different research contexts in which aggregation problems are endemic. These examples are intended both to be illustrative of the practical difficulties faced by researchers and to provide motivation for further abstract analysis of aggregation complications.

1. Grouping of Observations: A researcher has at his disposal (in principle) observations on the behavioral units of interest (e.g. persons, families, communities) but decides to summarize this information and employ grouped observations in the analysis. The analyst may decide to engage in this practice simply to reduce the magnitude of the analysis. Or, in a more interesting case, may be concerned with protecting the anonymity of the respondents.[2] For example, the researcher may feel obliged to guarantee anonymity to his respondents in a panel study and thus must use some "benign" identifying characteristics to compare early and later observations. For such an identifying characteristic to be innocuous it must identify only collections of individuals.[3] In such a situation the researcher will not be able to analyse relationships between individual responses but must employ the (grouped) observations on collectivities which are distinguishable over time. The most interesting feature of this case is that the researcher has control over the "grouping variable." He may choose characteristics like month of birth, father's occupational category, etc. The methodological and statistical problems which arise in situations like this have been discussed as issues of "grouping observations." (Prais and Aitchison, 1954; Cramer, 1964; Haitovsky, 1966).

2. Missing Data Problem: In analyses where a substantial portion of observations on key variables are missing and where the researcher has reason to believe that elimination of cases with missing data will systematically bias

results, it is important to attempt to estimate the missing data. While a number of procedures are available in the statistical literature, one more heuristic procedure found quite often in empirical research bears on our problem in an interesting way. In cases where there are no missing observations on one or more variables, it is possible to estimate missing values of other variables using information from the "completely measured" variables. For example, Kline, Kent and Davis (1971) in a cross-national analysis estimated the means of all partially measured variables for categories of nations grouped by date of independence, areal location, and political modernization (which were known [measured] for all nations).[4] All units with missing observations on other variables were then assigned the means for such variables for their category on the three "grouping variables." In other words, units are assigned the mean value of a variable for the category defined by values of completely measured variables. While this strategy raises a variety of measurement issues, what concerns us here is the dependence of the method on grouping.

3. Grouping to Minimize the Effects of Measurement Error: Blalock (1964) proposed the following strategy for handling random measurement error in independent (predetermined) variables. Search for an "instrument", i.e. a variable which affects the independent variable directly but which does not directly affect the dependent variable directly and is uncorrelated with excluded causes of the dependent variable. Then group observations according to this instrument and use grouped observations to infer the relationship of interest. More recent work by Blalock (Blalock, Wells and Carter, 1970) suggests that this may be generally less useful than an ungrouped instrumental variables approach or (in the face of specification error) than ordinary least squares. However, this work (along with the Wald-Bartlett methods) suggest that grouping observations may be one approach to resolving measurement difficulties. It is the consequen-

ces of grouping per se which we examine in this paper. But, it is obvious that success in overcoming measurement difficulties depends on the presence of "pure" grouping effects.

4. "The Aggregation Problem": The classic aggregation problem raised by economists concerns attempts to group observations on "behavioral units" so as to investigate economic relationships holding for sectors or total economies. The typical case involves merely macro-prediction. The aggregates are not usually conceded theoretical importance (thus the issues are not conceptual) but are deemed important for policy purposes. Since for many of the models employed there is no theoretical reason for suspecting that the processes holding at the level of firms, households, etc. would be different from those characterizing the behavior of more aggregated sectors, much interest focused on the conditions in which the inferences drawn from the relationships defined on the grouped observations would be consistent with those found with ungrouped observations. It is usually presumed in such discussions that the analyst has control over the grouping procedure. Theil (1954) has extensively discussed the aggregation complications which typically arise in this application. We will examine Theil's formulation below.

5. "Ecological Inference": Social scientists perhaps more frequently find themselves in the position of employing observations which were grouped for some other purpose or as the result of some social structural processes. These cases typically arise when we use observations grouped together by areal location (e.g. census tract means) or by location in some social structure (e.g. classroom or work-group means) or temporally grouped observations of frequent measurements (e.g. quarterly or yearly averages of monthly statistics). Much concern has focused on the consequences of using such data to make inferences to the relationship holding for the ungrouped observations. The crucial distinction here is

that the analyst does not have control over the grouping process and often does not understand the abstract consequences of the concrete grouping criterion employed (e.g. how does the census tract distinction coincide with neighborhood or social differences?) Sociologists and political scientists beginning with Robinson (1950) have discussed the problem as one of "ecological inference". Many of the issues were clarified by the application of linear models to the disaggregation problem by Goodman (1959) and by the application of a causal models perspective and an abstract conception of the effects of grouping by Blalock (1964).

The five analytic complications just outlined differ considerably in the intent of the analyst and in the necessity of relying on grouping of observations. Yet, it is clear that likely inference error produced by each strategy depends heavily on the existence (and likely magnitude) of "pure" grouping or aggregation effects. The formal similarities of the five situations will become clearer when we develop a formal model for assessing aggregation effects.

II AGGREGATION AND CONSISTENCY

In each case presented in Section I we have defined (whether or not it is directly available in an observational sense to the analyst) a micro-model expressed in terms of observations on micro-units. We have a set of grouping procedures or aggregation relations which define synthetic macro-observations as functions of micro-observations. Finally, we then consider a macro-model specified analagously to the micro-model (in terms of form of relation and variables included) defined on the macro-observations.

The three types of relationships (micro-model, macro-model, and aggregation relations) are not defined independently. When two of them have been specified, the third must take some limited form or the specification of relations will be internally inconsistent. This internal dependence of the system of relations

suggest a criterion by which to evaluate the effects of grouping observations. We will follow Green (1964) and define consistency to be the requirement that one be able to generate the same array of predicted macro-outcomes (dependent variables) by using the micro-model and aggregating predicted outcomes by the aggregation relation as by employing the macro-model directly. When this condition is satisfied we speak of consistent aggregation.

The consistency model is a very useful one for it allows the application of powerful mathematical analyses to the study of the conditions under which consistency is possible. Such study has had important consequences. A series of theorems by Leontief (1947a,b), Sono (1961) and Nataf (1948) prove that in the absence of very strong theoretical assumptions consistency is attainable only when all three relations are linear, even when the relations are deterministic.

Much of the literature in economics on aggregation problems attempts to develop highly specific theoretical models which result in consistency under specified conditions. An excellent review of much of this research has recently been done by Ijiri (1971). All of these attempts can be seen from the consistency perspective to involve one of three strategies: (1) fix the micro-model and the grouping relations and search for macro-models which result in consistency; (2) fix the macro-model and the "disaggregation" relations and search for a consistent micro-model; or (3) fix the micro and macro-models and search for consistent aggregation relations.

The methodological version of the aggregation problem involves a slightly different situation. We are interested in situations in which for practical purposes all three relations are fixed. The problem, then, turns on the use of an estimated macro-model to make inferences to the micro-model. We will always assume that the analyst implicitly formulates analagous micro and macro-models, i.e. models which have the same forms of relations and include the same varia-

bles. The restriction that they be analagous follows from the desire to substitute estimates from one model for the unavailable estimates from the more (or less) aggregated model. However, the practical situations differ in the degree of control that the analyst has over the grouping procedures. In the case where there is no control, it is obvious that all three relations are fixed. In cases where there is some control, we consider each of the alternative possibilities are distinct cases where all three relations are fixed. Then we proceed to evaluate the likelihood of erroneous inferences due to aggregation given relevant types of models and grouping procedures.

The mathematical analyses do demonstrate that the methodological problem as we have formulated it is generally intractable unless we limit our focus to cases where all three types of relations are linear, i.e. linear aggregation where both micro and macro-models are also linear. In addition, we largely limit our analysis to single equation (or recursive) micro and macro-models. The extension to just-identified micro and macro-systems is straightforward. However, the over-identified case has proven relatively intractable (Theil, 1959). Thus we can make no precise statements about the nature of aggregation effects in systems of interdependent equations.

III THE FORMAL MODEL

In what follows we will employ the following single equation micro-model:

$$y_i = \alpha + \beta_1 x_{1i} + \ldots\ldots + \beta_k x_{ki} + u_i \qquad (i=1,\ldots,N) \qquad (1)$$

where the x_{ij} are non-stochastic and the u_i have the usual good properties, i.e. $E(u_i u_i) = \sigma_u^2$, $E(u_i u_j) = 0$, and $E(u_i) = 0$ for all i,j. We may take as a substantive example a linear regression of pupil school achievement on pupil background characteristics (e.g. social class of parents, IQ) and educational and occupational aspirations and expectations.

It will be helpful to express the micro-regressions in matrix form:

$$y = X\beta + u \tag{2}$$

where y and u are N x 1 column vectors, and X is N x k with rank $k < N$. The restrictions on the population disturbances can now be expressed: $E(u'u) = \sigma_u^2 I$, $E(u) = 0$.

Next we consider a grouping relation defined on all N observations which takes arrays of micro-observations and substitutes the mean of the array as the observation. In the substantive example, this might involve the grouping of pupil observations into classroom or grade-level means. Abstractly, we consider the application of a grouping matrix G (m x N) with $m < N$ (where m is the number of groups) such that

$$\bar{y} = Gy \qquad \bar{X} = GX \qquad \bar{u} = Gu \tag{3}$$

For example, if the first group includes the first n_1 observations, the second group the second n_2, etc.:

$$G = \begin{bmatrix} 1/n_1 \ldots 1/n_1 & 0 \ldots 0 \\ 0 \ldots 0 & 1/n_2 \ldots 1/n_2 \; 0 \ldots 0 \\ \vdots & \\ 0 \ldots 0 & 1/n_m \ldots 1/n_m \end{bmatrix}$$

We will usually consider the case where $n_i = n_j$ for all i,j, i.e. equal sized groups. We will see below that deviation from equal sized groups creates estimation problems.

It is important to note that the matrix G merely summarizes the consequences of the application of a grouping rule. The grouping rule specifies which elements in G are non-zero, i.e. determines which observations are to be consolidated. We will be preoccupied with the underlying logic of the grouping proce-

dure -- what we are calling the aggregation relation.

Finally, we define a macro-model which is analagous to the micro-model:

$$\bar{y} = \bar{X}\bar{\beta} + \bar{u} \quad . \tag{4}$$

where $\bar{y}$ (m x 1), $\bar{X}$ (m x k), and $\bar{u}$ (m x 1) are given by (3). Since this model is not specified independently of (2) and (3) we cannot specify its properties in the abstract. As we shall see, the substance of an aggregation analysis consists precisely in determining the dependence of the properties of (4), (particularly those relating to the behavior of the disturbance vector, $\bar{u}$), on (2) and (3).

Since this model allows for the grouping of observations but not of variables, the coefficient vector, $\bar{\beta}$, of the macro-model has the same order as that for the micro-model. Thus every coefficient in one model has a corresponding term in the other. We do not necessarily presume, however, that (hypothetically) estimated micro and macro-coefficients are produced by the same estimation procedure. We will tend to restrict our attention to the case where ordinary least squares regressions (OLS) is applied to each. But, we will see below that there are cases where this is not the optimal approach.

The last step in specifying the model is to apply the consistency criterion to this three-relation system. We noted above that consistency can be seen to require that one generate the same values of the macro-dependent variable using (4) directly as by using (2) and then applying (3) to the generated micro-values. This requirement is overly strong for the cases we are considering and we will relax it to require only that the expected values of the predicted $\bar{y}$ be the same under each method. We will call this weaker version, <u>stochastic consistency</u>. For the system of relationships outlined above, the macro-procedure for generating $\bar{y}$ values uses simply $\bar{X}\bar{b}$ (where $\bar{b}$ is a vector of estimates of $\bar{\beta}$). The micro-approach employs Xb (where b is a vector of estimates of β) and

then applies G to $\hat{y} = Xb$. Thus to assess consistency we compare $\overline{Xb}$ and GXb. Since by (3) $\bar{X} = GX$, the consistency criterion for the class of aggregation situations we are considering is that $E(\bar{b}) = E(b)$.

This argument should be modified in one small detail. We must consider the possibility that estimates from the micro-model are biased, i.e. $E(b) \neq \beta$.[5] The practical situations we are addressing can be seen to involve the substitution of the estimated macro-coefficients for the unknown micro population parameters. Thus the micro-estimates are not unambiguous guides for evaluating the consequences of employing the macro-estimates. It seems more reasonable to argue that stochastic consistency requires that

$$E(\bar{b}) = \beta. \tag{5}$$

We will employ this version in analyzing aggregation difficulties.

When aggregation is inconsistent in simpler linear models of the sort we are considering, we speak of aggregation bias. We can define such bias simply using (5): aggregation bias is

$$E(\bar{b}) - \beta \ . \tag{6}$$

Obviously, the situation is simplest when the estimates of the micro-coefficients are unbiased. For then, we can use a simpler expression for aggregation bias:

$$E(\bar{b}) - E(b) \ .$$

Most methodological treatments of the aggregation problem have concentrated exclusively on bias in correlation and regression coefficients. Sociologists, in fact, rarely devote any attention to other properties of estimators. Yet, as is demonstrated in any introductory statistical inference text, there are reasonable situations where biased estimators but small variance are preferred to less efficient unbiased estimators. Since all aggregation involves some loss of information, we should expect the efficiency of estimators to be affected. Thus we will broaden the conventional sociological focus at least to the point

of addressing the consequences of aggregation for the efficiency of estimators.

IV THREE GROUPING PERSPECTIVES (THE BIVARIATE CASE)

In this section, we will somewhat arbitrarily categorize a number of approaches to the grouping problem into three perspectives: (1) a clustering perspective focusing on the grouping of "natural" units; (2) an optimal grouping approach which presumes that the analyst has control over the aggregation process; and (3) a causal models perspective. As we shall see, although these perspectives share a somewhat common focus, they are not equally useful and suggestive of analysis problems. And, at the same time, these perspectives do not exhaust the subject matter. In the following sections we will develop some alternative formulations which partially complement the dominant findings reported in this section, and allow treatment of multivariate models.

(1) The Clustering Perspective. Apparently the earliest concerns with aggregation problems in the social sciences arose over the inflation of correlation coefficients as units of observation were grouped together. This effect was noticed in a wide variety of applications(e.g. correlation of rental values and delinquency rates for city subareas (Gehkle and Biehel, 1934), correlation of crop yields for different crops in regions (Yule and Kendall, 1950), correlation of race and literacy in the United States, (Robinson, 1950). In each case the increase in linear correlation was thought to be artificial and attempts were made to uncover the mechanism responsible for the artifact. A number of different algebraic formulations lead largely to the same account. We will briefly outline the approach which decomposes analysis of variance formulae since it provides the clearest demonstration of the inflation mechanism.

We will employ a hybrid notation to clarify the relations of this literature to the one considered in the next section. We denote samples variances

and covariances as follows.

$$C_{xx} = \frac{1}{N} \sum_{r=1}^{R} \sum_{i=1}^{N} (x_{ir} - x..)^2$$

$$C_{xy} = \frac{1}{N} \sum_{r=1}^{R} \sum_{i=1}^{N} (x_{ir} - x..)\,(y_{ir} - y..) \ .$$

The within group variances and covariances are denoted

$$WC_{xx} = \frac{1}{N} \sum_{r=1}^{R} \sum_{i=1}^{N} (x_{ir} - x._r)^2$$

$$WC_{xy} = \frac{1}{N} \sum_{r=1}^{R} \sum_{i=1}^{N} (x_{ir} - x._r)\,(y_{ir} - y._r)$$

and the between-group ("ecological") variances and covariances:

$$C_{\overline{xx}} = \frac{1}{R} \sum_{r=1}^{R} (x._r - x..)^2$$

$$C_{\overline{xy}} = \frac{1}{R} \sum_{r=1}^{R} (x._r - x..)\,(y._r - y..) \ .$$

This notation assumes there are N micro-units and R groups.

In the sociological literature the problem we are now addressing has continually been referred to (following Robinson's (1950) designation) as the "ecological correlation" problem. To simply the algebra we specialize the micro-model developed in Section II to include only one regressor. In this case it is easy to prove:

$$C_{xy} = WC_{xy} + C_{\overline{xy}} \quad ;$$

and thus:

$$R_{xy} = WC_{xy} + C_{\overline{xy}} \Big/ \sqrt{C_{xx} C_{yy}} \quad ;$$

or

$$R_{xy} = WR_{xy} \sqrt{1-E_{yr}^2}\ \sqrt{1-E_{xr}^2} + R_{\overline{xy}} E_{yr} E_{xr}$$

where E_{xr} and E_{yr} are the "correlation ratios" for grouping or region (where

$$E^2_{xr} = \frac{C_{\overline{xx}}}{C_{xx}} \Bigg) \quad .$$

The translation from micro-correlations (R_{xy}) to "ecological correlations" ($R_{\overline{xy}}$) is not simple because the within-group or within-region correlations are not simple arithmetic means of the micro-correlations. However, following Robinson's argument we can see that two things typically occur as micro-units are consolidated:

1. The within-group correlation WR_{xy} increases due to increasing heterogeneity of groups and this effect decreases the ecological correlation since the proportion of the variance "explained" by the grouped observations is equal to $1-WR^2_{xy}$.
2. The values of the correlation ratios E^2_{xr} and E^2_{yr} decreases as a consequence of the decreased variability of X and Y values in the grouped observations.

But, Robinson (1950, pp. 356-357) argued:

> ... these two tendencies are of unequal importance. Investigation of (3.11) with respect to the effect of changes in the values of E^2_{xr}, E^2_{yr}, and WR_{xy} indicates that the influence of the changes in the E's is considerably more important than the influence of changes in the value of WR_{xy}. The net effect of changes in the E's and WR_{xy} taken together is to increase the numerical value of the ecological correlation as consolidation takes place.

This argument is demonstrably not universally true and depends on an unstated assumption which remained implicit in much of the thinking of sociologists prior to the exposition of the causal models approach discussed below. As we will see, if observations are grouped randomly, the expected value of the sample ecological correlation coefficient is equal to the expected value micro-correlation. But, the grouping relation assumed by Robinson is never clearly specified beyond the assertion that with grouping relatively more homogeneous units are consolidated into more heterogeneous units. We are, however, to assume that the groups are formed on the basis of administrative units, e.g.

census tract boundaries.

In the case of census tracts we know that the boundaries were devised to correspond as closely as possible to "natural areas." This suggests that using data grouped by census tracts will combine observations on units which are relatively homogeneous on a great variety of sociological variables, e.g. race and literacy. Thus areal grouping in such a case is not random with respect to the variables included in the regression but may be thought to systematically affect their variation. Such a situation will have the consequences Robinson described. But, it is important to note that the inflation mechanism depends on the substantive assumption of a nonrandom areal distribution of properties which correspond in some way to the administrative boundaries. In fact, as we shall see, the more nonrandom the distribution and the closer the correspondence with the grouping boundaries, the greater the inflation of the ecological correlation over the micro-correlation.

Robinson created something of an intellectual controversy by asserting the ecological correlations are always computed for some more micro-interest, and, further, that ecological correlations never provide useful information about micro-relationships. The first argument need not be answered and the second stimulated several innovative attempts at applying linear models (Duncan and Davis, 1953), Goodman (1953, 1959) and non-linear models (Boudon, 1963) to the problem of using ecological correlations to estimate or to set limits on the micro-correlation. These proposals, which have never achieved widespread use in sociology (even though the use of ecological correlations for micro-pursuits continues almost unabated), are treated in some detail by Alker (1969), Stokes (1969), and Hannan (1971), and will not be discussed here.

(2) Optimal Grouping. Economists have frequently faced difficulty in dealingwith an overabundance of data on households and have considered the conse-

quences of summarizing the data by alternative grouping procedures. Prais and Aitchison (1954) and Cramer (1964) have provided the basic results. These analyses treat the practical problems facing the investigator contemplating alternative groupings, thus the methodological development presumes control over the definition of the grouping criterion.

This analysis continues to use the two-variable model developed above. But this time we employ sums of squares and crossproducts rather than variances and covariances. Thus we define:

$$S_{xx} = \sum_{r=1}^{R} \sum_{i=1}^{N} (x_{ir} - x..)^2 \tag{7}$$

$$BS_{xx} = \sum_{r=1}^{R} N_r (X_r - X..)^2 \tag{8}$$

$$S_{xu} = \sum_{r=1}^{R} \sum_{i=1}^{N} (x_{ir} - x..)(u_{ir} - u..) \tag{9}$$

$$BS_{xx} = \sum_{r=1}^{R} N_r (x._r - x..)(u._r - u..) \tag{10}$$

with within-group sums of squares similarly defined. Substitution (7) and (8) into the micro-model yields:

$$S_{xy} = \beta S_{xx} + S_{xu}$$

and $$BS_{xy} = \beta BS_{xx} + BS_{xu} .$$

The regression coefficient for the micro-model is given by:

$$b = \frac{S_{xy}}{S_{xx}} = \beta + \frac{S_{xu}}{S_{xx}}$$

which has expected value:

$$E(b) = \beta .$$

The regression coefficient for the macro-model is given by

$$\bar{b} = \frac{BS_{xy}}{BS_{xx}} = \beta + \frac{BS_{xu}}{BS_{xx}}$$

which also has expected value:

$$E(\bar{b}) = \beta \quad .$$

Thus for the model as specified, there is no aggregation bias. Obviously, this important result depends on the fact that $E(BS_{xu}) = 0$. As we shall see in the discussion of the causal models approach, there is an important class of G where $E(S_{xu}) = 0$ while $E(BS_{xu}) \neq 0$. Clearly such a case violates the result just cited.

As was mentioned in passing at the end of Section III, we should expect grouping to result in the reduction of efficiency of the macro-estimators. To see this, we compare the variances of the two estimators:

$$\text{var}(b) = \frac{\sigma_u^2}{S_{xx}} \qquad \text{and} \qquad \text{var}(\bar{b}) = \frac{\sigma_u^2}{BS_{xx}}$$

Thus the efficiency of estimation of the macro-coefficient in the grouping specified above is given by BS_{xx}/S_{xx}. This last term is necessarily less than or equal to unity since

$$S_{xx} = WS_{xx} + BS_{xx}$$

where in the model as specified all three terms are nonnegative and $\text{Cov}(WS_{xx}, BS_{xx}) = 0$. Thus for the class of aggregation relations considered, aggregation reduces efficiency.

The efficiency result suggests a practical research strategy. Since the closer BS_{xx} is to S_{xx}, the less the loss of efficiency, an analyst who has control over the grouping procedure ought to choose a G which maximizes variation in X. This issue is quite important since efficiency can be greatly reduced in common practice. For example, Cramer demonstrates that for random grouping N micro-observations into m groups, the loss of efficiency is approximately m-1/N-1. Thus for example when, say, 500 observations are randomly placed in

25 groups, the efficiency of the grouped estimator is approximately .048 relative to the micro-estimator.

Cramer (1964) has shown in addition that, holding the number of micro-observations constant, efficiency of the macro-estimator decreases with the "coarseness" of the grouping procedure. That is, the greater the level of consolidation (the greater the loss of information) the lower the efficiency. Thus we find the expected tradeoff between economy (small number of groups) and efficiency (small variance).

Another potential difficulty presents itself when the groups are of unequal size. Consider equation (4), the original macro-model. From the properties of the disturbance vector, it is clear that

$$E(\bar{u}) = 0 \quad \text{and} \quad E(\bar{u}\,\bar{u}') = \sigma_u^2 G'G \quad .$$

Thus the disturbance term for the macro-model does not have the diagonal form which makes ordinary least squares a "best" estimator. The disturbance term $\sigma^2 GG'$ is heteroscedastic. Following Prais and Aitchison (1954) we see that the best unbiased linear estimator for the macro-model is Aitken's generalized least squares (GLS). The GLS estimator of β from the macro-model is

$$b = [\bar{X}'(GG')^{-1}\bar{X}]^{-1}\,\bar{X}'(GG')^{-1}\bar{y}$$

with variance-covariance matrix

$$\text{var}(b) = \sigma^2[\bar{X}'(GG')^{-1}\bar{X}]^{-1}$$

The generalized variance term GG' takes on a simple form for the matrix defined in (3). It is an m x m matrix:

$$GG' = \begin{bmatrix} 1/n_1 & 0 & \cdots & 0 \\ 0 & 1/n_2 & \cdots & 0 \\ \vdots & & & \\ 0 & \cdots & & 1/n_m \end{bmatrix}$$

so that

$$(GG')^{-1} = \begin{bmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & & & \\ 0 & \cdots & \cdots & n_m \end{bmatrix} .$$

Thus the application of GLS involves weighting grouped observations by the number of micro-observations used to construct the group. And, it is clear that when the number of observations comprising each group is the same (i.e. $n_i = n_j$ for all i,j) ordinary least squares is the GLS estimator.

Using this formulation it is easy to show how the inflation of correlation coefficients depends on the type of grouping procedure. It is most convenient to assume that the m groups are of equal size, say n_i and to use the following expression for the estimated micro-correlation:

$$R^2_{xy} = 1 - \frac{mn_i\sigma^2_u}{\sum_{ij}(Y_{ij}-Y..)^2} = 1 - \frac{mn_i\sigma^2_u}{S_{yy}}$$

The macro-correlation is given by:

$$R^2_{\overline{xy}} = 1 - \frac{m\sigma^2_{\overline{u}}}{\sum_j n_i(Y_{i.} - Y..)^2}$$

With linear aggregation of the type under consideration $\sigma^2_{\overline{u}} = \sigma^2_u/n_i$.

Thus macro-correlation can be written:

$$R^2_{\overline{xy}} = 1 - \frac{m\sigma^2_u}{BS_{yy}} .$$

Given the basic theorem from the analysis of variance:

$$S_{yy} = WS_{yy} + BS_{yy} ,$$

it follows that if the grouping procedure is random,

$$S_{yy}/(mn_i-1) \text{ and } BS_{yy}/(m-1)$$

are unbaised estimates of the same variance. Thus we should expect that $R^2_{\overline{xy}}$ will be close in value to R^2_{xy}. Cramer's (1964) analysis demonstrates this is the case.

But, what about grouping which maximizes variation in X? It is clear in that case that BS_{yy} will be larger than in the random grouping case. As BS_{yy} increases over the value it would take on in random grouping the macro-correlation exceeds the micro-correlation.

(3) A Causal Models Approach. The previous section makes plain the fact that the consequences of grouping or aggregation depends on how G affects variation in the variables of the model under study. We have seen that random grouping does not produce aggregation bias in either correlation or regression coefficients. Grouping which maximizes variation in a regressor in a bivariate model produces aggregation bias in the correlation but not in the regression coefficient. But, this type of grouping is more efficient in the statistical sense than random grouping. But, what about grouping which maximizes variation in the dependent variable or regressand?

Blalock (1964) considered, from both a formal and a causal point of view, the case where variation in the regressand is maximized by the aggregation rule. From a formal perspective it is clear that the correlation coefficient, $R^2_{\overline{xy}}$ behaves symmetrically with respect to changes in variation of either X or Y. Thus grouping which maximizes variation in either variable will inflate (bias) the macro-correlation coefficient. But, $R^2_{\overline{xy}} = b_{\overline{yx}}b_{\overline{xy}}$ by definition and if grouping by X increases $R^2_{\overline{xy}}$ and leaves $b_{\overline{yx}}$ unchanged, the $b_{\overline{xy}}$ must be increased proportionately to $R^2_{\overline{xy}}$. The slope $b_{\overline{xy}}$ is the "wrong" slope from the point of view of the micro-model specified in (2). And, its bias can be considered a mathematical artifact. But, what of the case where variation in Y is maximized where the substantive interest is in model (2)? It is clear that with such

grouping $b_{\overline{xy}}$ the "wrong" slope will be unbiased and the slope of interest will be biased.

We gain additional insight into the mechanism operating to produce the bias when we approach the problem from a causal models perspective. To maximize variation in Y we rank observations by Y values and then apply a G which groups "adjacent" observations together. Since by (2) Y is a linear function of both X and U, such a G will place in the highest Y groups observations which have both high X and high U values and similarly observations with both low U values for the groups lowest on Y. This G confounds X with other causes of Y. This confounding of the variation in X and U with respect to variation in Y has very serious consequences. While in the micro-specification $\sigma^2_{xu} \xrightarrow[\text{plim}]{} 0$ this will no longer be the case for the macro-model. The correlation in probability limit of the disturbances and independent variables violates the specification legitimating OLS and is usually called a specification error. Since the macro-model is misspecified in the case of grouping by Y we would not expect OLS to have desirable properties. This is the case since OLS are now biased and it is this bias which we term aggregation bias.

It is a simple extension to argue that grouping by any endogenous variable in a simultaneous system will tend to produce aggregation bias in macro-estimates. This should be the case since the grouping mechanism specified will produce covariation between disturbances and exogenous variables in the equations for the endogenous variables which are systematically grouped. Obviously it will be more difficult in practice to ascertain before the fact the consequences of systematic grouping of observations in a simultaneous system of equations.

We can gain some understanding of the likely direction and magnitude of the aggregation bias from a knowledge of the micro-model's properties (although in

the practical situations outlined at the outset this information will not always be available). Shively (1969) and Hannan (1971) show for a discontinuous and continuous model respectivity that σ_{xu} will be positive. And, as a result, the direction of the aggregation bias from grouping by Y will be in the direction of b_{yx}. Thus when b_{yx} is positive, grouping b_y will inflate b_{yx}. Secondly, we can identify a condition under which the likely aggregation bias will be small. Shively (1969) has suggested that when R^2_{xy} is large in magnitude, grouping by Y will not produce large bias. We can see that the stronger the linear association of X and Y, the less important will be the causes of Y which are confounded with X and the more grouping by Y will approximate grouping by X (which produces no bias). That is, U values will become less important in determining the ranking of any observation on Y relative to the U value. Thus we should expect σ_{xu} to decline as R^2_{xy} approaches unity. Thus the less causally important relative to X are the factors ignored in the micro-model, the lower the aggregation bias from grouping by the dependent variable.

The reader who has considered the aggregation problem only for cases where the analyst has control over the grouping operation may question the practical importance of this case. After all, why would someone group observations by the dependent variable? The answer is that when data is aggregated "naturally" i.e. as a consequence of social structural process such as bureaucratic administration, the grouping procedures often do not operate explicitly on some variable but on some concrete property of micro-units. The property in question is almost always "location" in some social structural space: a residence, classroom in school, etc. The practical aggregation problem of grouping observations by the dependent variable arises because of the possibility that location in the particular social structural space which is utilized in the grouping corresponds to variation in the variable which the analyst wishes to take as

dependent. Consider the educational achievement model outlined above. It is possible that administrative policies may locate students of the same age in classrooms according to measured academic performance. If data gathered on students in the schools in such a system is aggregated by classroom, the consequences of using the grouped data obviously depend on the model under study. If, as in the example, achievement is taken as dependent in one equation in the model, then the model estimated from the grouped data will contain aggregation bias. This example suggests broad classes of situations in which grouping by the dependent variable may occur.

The discussion of grouping effects like the literature it follows focused wholly on the bivariate cases. As we noted above, additional complications arise in multivariate single equation models. It is to this issue that we now turn. We will consider first a classic formulation of the problem by Theil (1954) and then move on to develop a more general specification error argument.

Before doing this, it is useful to summarize the results for the bivariate case. The effects of grouping on correlation and regression coefficients estimated from the macro-model depend on the nature of the grouping procedure. We considered three possibilities: (1) random grouping, (2) grouping by X, and (3) grouping by Y. Random grouping does not produce any aggregation bias but does reduce efficiency of both correlation and regression estimators. Grouping by X is also unbiased with respect to regression estimators and is more efficient than random grouping. It does, however, inflate (bias) correlation coefficients. Finally, grouping by Y biases both correlation and regression coefficients for the model in which Y is taken as dependent. The magnitude of the bias which will be positive for the correlation coefficient and of the same sign as the regression coefficient, decreases as R^2_{xy} decreases.

V AGGREGATION BIAS AND SPECIFICATION ERROR

1. Specification Error: In statistical analysis we commonly deal with the empirical consequences of some "maintained hypotheses" or statistical model. It is typical not to question the truth status of the entire model but to doubt only certain specified elements ("null hypotheses"). If the model is inconsistent with evidence in such an analysis all of the blame is attached to the specified null hypotheses. Thus the micro-model (2) specified above is tacitly accepted to be a good representation of reality and one might hypothesize, for example, that some coefficient is zero in a population of interest, and test this hypothesis with a sample of observations. But, in applied work we seldom completely trust our models. That is, there might be competing substantive arguments which claim that true model is nonlinear, or contains additional variables, etc. It is important, then, to consider the consequences of conducting statistical analyses on fallible models. Following Theil (1957), the problem can be formulated as follows. We proceed with some model $\bar{M}$ yielding estimates $\bar{\theta}$ which are thought to have good properties when the true model is M which yields estimates θ. The methodological question is: what can be said about the estimates $\bar{\theta}$ in light of the knowledge that M is the correct model.

Of course, there are a great many possible departures of models from the true models whose consequences are not easily established formally. However, we can deal explicitly with some classes of departures, which we call specification errors. For the study of aggregation complications two types are most interesting: (1) the use of the "wrong" variables in an otherwise good model, i.e. observations from the wrong level of aggregation: and (2) the exclusion of causally important variables from the model. We will phrase the aggregation problem in terms which correspond to each case in turn.

We have to transform our basic problem temporarily to make use of the specification apparatus developed by Theil. We continue to accept the micro-model

(2) as the true model. However, we now consider an alternative specification (analagous to the macro-model on the right hand side with the left hand side unchanged from (2)):

$$\bar{y} = \bar{X}\beta + \bar{u} \quad . \tag{11}$$

Given the specification $E(u) = 0$, $E(Xu) = 0$, $E(u'u) = \sigma^2 I$

$$b = (X'X)^{-1}X'Y \quad . \tag{12}$$

is the test estimate of β in (2). Since we are considering <u>nonstochastic</u> regressors, it also follows that the OLS estimate of $\bar{\beta}$ in (11) is

$$\bar{b} = (\bar{X}'\bar{X})^{-1}\bar{X}'Y \quad . \tag{13}$$

The point of this departure is to comment on the use of $\bar{b}$ for purposes of making inferences about β. We can use the following important theorem.

THEOREM: Suppose the micro-model (2) is true and that X is some matrix with real, nonstochastic elements. Then the statistic $\bar{b}$ of (13) is an unbiased estimate of

$$P\beta \ (= E\bar{b}) \tag{14}$$

where P is the coefficient matrix of the least-squares regressions of X (the correct explanatory variables) on $\bar{X}$ (the incorrect ones):

$$P = (\bar{X}'\bar{X})^{-1}\bar{X}'X \quad . \tag{15}$$

The proof is straightforward:

$$E\bar{b} = (\bar{X}'\bar{X})^{-1}\bar{X}'Ey = (\bar{X}'\bar{X})^{-1}\bar{X}'X\beta = P\beta \quad . \tag{16}$$

Theil uses the term auxiliary regressions to refer to P. These regressions take the form

$$X = \bar{X}P + \text{matrix of residuals.}$$

To see how they are employed, consider a simple case of misspecification. Suppose that $\bar{X}$ is identical to X in all but the last column where the true variable X_k is replaced by some X_k'. The matrix P will be a unit matrix except for the last column which will include all non-zero entries. And, we see that in general, each element of $E\bar{b}$ depends not only on the corresponding β but on the

β-component of the incorrectly specified variable X_k. That is

$$E\bar{b}_j = \beta_j + P_{jk}\beta_k \qquad j = (1,\ldots,k) \tag{17}$$

where the p's are the coefficients of the auxiliary regression.

$$X_k(t) = \sum_{j=1}^{k-1} P_{jk}X_j(t) + P_{kk}X'_k(t) + E \quad .$$

The excess term in (17), $E\bar{b}_j - \beta_j$ is called the <u>specification bias</u> of (11). (The parallels with our formulation of aggregation bias are obvious.) It is important to note that if other explanatory variables are uncorrelated with the <u>correct</u> regressor, their coefficients do not contain any aggregation bias.

The above example treats the inclusion of a "wrong" variable for a "correct" one. One such case which is frequently encountered in practice is the substitution of an imperfectly measured indicator for the propert("true") causal variable. Another type of specification error which will be useful in the treatment of aggregation problems is the exclusion of a causal variable, i.e. the omission of a variable which "belongs" in the model. The specification bias for the latter case takes the same general form. It will be useful for later analysis to work through a three variable example. Let the correct specification be

$$y_i = \beta_{y1}x_{1i} + \beta_{y2}x_{2i} + u_i \qquad (i = 1,\ldots,N) \tag{18}$$

and the estimated incorrect specification be

$$y_i = b_{y1}x_{1i} + e_i \quad . \tag{19}$$

The specification bias in (19) is given by:

$$E(b_{y1}) = P\beta$$

where β is $\begin{pmatrix} \beta_{y1} \\ \beta_{y2} \end{pmatrix}$. As above P is given by:

$$P = (\bar{X}'\bar{X})^{-1}\bar{X}'X$$

where X is $\begin{pmatrix} x_{11} & x_{21} \\ \vdots & \vdots \\ x_{1n} & x_{2n} \end{pmatrix}$ and $\bar{X}$ is $\begin{pmatrix} x_{11} \\ \vdots \\ x_{1n} \end{pmatrix}$.

Working through the algebra yields:

$$P = \left(\frac{\Sigma x_{11}^2}{\Sigma x_{11}^2} , \frac{\Sigma x_{1i}x_{2i}}{\Sigma_{x1i}^2} \right) = (1, b_{21})$$

where b_{21} is the auxiliary regression of X_2 on X_1, and

$$E(b_{y1}) = P\beta = (1, b_{21}) \begin{pmatrix} \beta_{y1} \\ \beta_{y2} \end{pmatrix}$$

$$= \beta_{y1} + b_{21}\beta_{y2} \quad . \tag{20}$$

Thus, the expected value of the estimated X_1 coefficient in the misspecified model is equal to the corresponding population parameter for the correct model plus a specification error term. We see that the specification error term is nonzero under the condition that X_1 and X_2 are correlated in the sample and that X_2 has an "independent" linear effect on Y_1.

We now turn to the aggregation implications of the two cases just discussed: (1) the substitution of the wrong variables in the model; and (2) the exclusion of causal variables from the model. The application of the first case brings us to a consideration of a distinctive formulation of aggregation problems due to Theil (1954). Theil's work proceeds independently of discussion of various types of grouping problems.[6] And, for second case brings us directly back to approaches used to deal with the complications introduced by grouping of observations. The economy of the specification error approach is clearly indicated by the demonstration of the formal similarities of two quite different methodological traditions. In fact, we will ultimately see that we

can straightforwardly translate most of what is known about aggregation bias into this framework without sacrificing any essential detail.

2. Theil's Formulation of Aggregation Bias

In treatments of what we called areal aggregation the homogeneity-heterogeneity of micro-units is quite important. As we have seen, if areal units are more homogenous than are more inclusive areal groupings, then areal aggregation will not be random grouping but will be grouping which systematically affects variation in many substantive variables. For this reason the analysis of areal grouping necessarily involves the study of social homogeneity in spatial distributions. The notion of homogeneity employed refers simply to possession of some property, e.g. income. Thus income-homogeneous micro-units are those which have incomes in the same arbitrarily defined categories. In these terms, areal units which have identical income distributions are homogeneous

To appreciate Theil's approach we need to extend the homogeneity notion somewhat. A more fundamental conception of homogeneity would seem to include the idea that units (no matter what the magnitude of the properties of interest they possess) behave alike with respect to changes in causal variables. In this sense, micro-units in different income categories would be considered inhomogeneous only in the case that they react differently (in terms of some other variable) to a unit change in income. In the linear models framework we are using, the issue turns on variability in regression parameters. Since we almost always employ cross-sectional models, we must assume no inter-unit variability in regression parameters. Thus as a consequence of our analysis models, we implicitly hold to the notion of homogeneity just raised. The only language we have for dealing with departures from homogeneity is to speak of interaction effects. In this language we say that the way in which units react to changes

in income depends on some other property of the units,e.g. race, culture, etc.

When we have available panel observations with long time series, we can estimate the causal parameters which describe the behavior of each micro-unit. In any empirical analysis we would expect that sampling error alone would produce variability in estimated parameters for different micro-units. But, we may argue that there are other (systematic or random) factors which account for this inter-unit variability in response. In any concrete analysis the nature of the factors producing heterogeneity will be very important in determining the consequences of various complications. Here, we merely wish to demonstrate the likely aggregation consequences of this sort of heterogeneity. After we have developed the formal model we will return to the issue of types of sources of heterogeneity.

It will be simplest to modify our formal model slightly. Let the micro-model be

$$y_i(t) = \beta_{1i}x_{1i}(t) + \ldots + \beta_{ki}x_{ki}(t) + u_i(t) \qquad \begin{matrix}(t = 1,\ldots,T)\\(i = 1,\ldots,N)\end{matrix} \tag{21}$$

where we allow each micro-unit to have its own set of causal parameters. Note that we have T observations on each of the N micro-units. As previously, we define group relations

$$\begin{aligned} y(t) &= \frac{1}{N_i}\Sigma_{yi}(t) \\ \bar{x}_k(t) &= \frac{1}{N_i}\Sigma x_{ki}(t) \end{aligned} \tag{22}$$

The macro-model employed in the regression analysis is

$$y(t) = \bar{b}_1\bar{x}_1(t) + \ldots + \bar{b}_k\bar{x}_k(t) + u(t) \text{ or } y = \overline{xb} + u \tag{23}$$

where $\bar{b}$ are OLS estimators, i.e. $\bar{b} = (\bar{X}'\bar{X})^{-1}\bar{X}'y$.

To make use of the specification error approach we notice that if (21) is the correct micro-specification, then given (22), the grouping relations, the

macro-specification should not be (23) but

$$y(t) = \beta_1^* x_1(t) + \ldots + \beta_k^* x_k(t) + u^*(t) \qquad (24)$$

$$\text{or} \quad y = \bar{x}\beta^* + \underline{u}^*$$

where

$$\beta_k^* = \frac{1}{N} \sum_i \beta_{ki} \quad ,$$

the mean of the coefficients for all micro-units. Thus, for this type of grouping, aggregation bias is defined as

$$E(\bar{\beta}_k) - \beta_k^* \quad . \qquad (k = 1,\ldots,k)$$

Since (24) is the correct specification, the use of (23) in the regression should result in bias. Applying Theil's theorem, we obtain

$$E(\bar{b}) = P\beta^* \quad ,$$

where P is the matrix of auxiliary regressions

$$P = (\bar{X}'\bar{X})^{-1}\bar{X}'X.$$

From the earlier discussion it is clear that unless each micro-variable is uncorrelated with all "non-corresponding macro-variables", i.e. unless P is an identity matrix, we have aggregation bias. This means that estimated macro-coefficients depend generally on the parameters associated with noncorresponding micro-variables. In the educational achievement example this would mean that, unless micro-units were homogeneous regression estimates from a time series of grouped observations the effect of IQ on achievement, say, would depend not only on the ways in which micro-units IQ's determined achievement; but this term would also depend on, for example, the parameters relating aspirations to achievement.

Theil has made clear the fact that as long as there is a linear partial association between noncorresponding micro and macro-variables, aggregation bias (from the micro-perspective) will result from the regression analysis of

macro-specification. Boot and deWit (1960) have demonstrated the existence of such bias in a simple production function analysis.

It is somewhat difficult, however, to imagine how a practicing social scientist would determine whether or not nonzero partial associations between noncorresponding micro and macro-variables are likely in his analysis. In previous work (Hannan 1971), I have suggested developing 'cross-level models" to specify the effects over time of noncorresponding macro-variables on micro-variables. In the achievement example, such an effect might involve changes in individual aspirations as a consequence of changing social class composition of the classroom. This sort of thinking, sometimes called "structural effects" or "compositional effects" modeling, is the subject of considerable methodological and metamethodological debate in sociology at the current time. It is obvious, despite the virtues of this thinking from any substantive perspective, that the existence of such cross-level effects would produce aggregation bias in the type of model we are considering.

The situation when formulated in these terms becomes somewhat more complicated. If noncorresponding macro-variables have causal effects on micro-variables, we are tempted to argue that the original micro-model is misspecified. Grunfeld and Griliches (1960) develop a number of interesting aggregation implications of situations of that sort -- where the micro-model should contain the macro-variable. But, the type of cross-level effect we are discussing has a different status. Notice that specification bias requires both collinearity of included and excluded regressors and an independent causal effect of excluded variables on the dependent variable. Thus a classic specification error formulation would seem to require that noncorresponding macro-variables not only effect micro-variables but have independent causal effects on the micro-dependent variable. Theil's result does not require the independent causal

effect. Aggregation bias will obtain so long as noncorresponding macro and micro variables are causally related whether or not any macro-variables "belong" in the micro-specification. To return to the substantive example, this means that aggregation bias would result from the estimation of the macro-relation if individual aspirations had a nonzero partial relation (in the auxiliary regression) with mean social class even if mean social class has no effect on individual achievement. This, then, is an extremely important result since it is considerably stronger than the specification error result.

A second feature of Theil's formulation worthy of notice is that it does not specify anything about the nature of the grouping relations other than their linearity. The connection between the nature of the grouping and this form of the aggregation bias has never been made clear. It appear, however, that the nature of the grouping relation to a large extent determines the behavior of the auxiliary regressions. One reason why it is so difficult to specify anything about the behavior of the auxiliary regressions in the abstract is that nothing has been assumed about the nature of the grouping operation. It seems likely that Theil has randomgrouping in mind for we shall see in a later section that grouping by one of the regressors will produce a different type of specification error. And, the earlier result on grouping by Y holds with equal force here.

Random grouping does not appear to result in auxiliary regressions with nonzero coefficients associated with noncorresponding macro-variables.[7] The earlier discussion of random grouping would seem easily generalized to the multivariate panel case under consideration here. Thus, we see that it is important to deal more explicitly with grouping by micro-regressors. It is to this subject that we now turn.

3. Grouping and Specification Error

Here we take up the second class of specification error problems outlined in the formal discussion: the exclusion of a collinear causal variable from the micro-model. This analysis depends completely on the work of Hiatovsky (1966). The discussion will be simplified if we refer to a three-variable micro-model:

$$y = \beta_1 x_1 + \beta_2 x_2 + u \tag{25}$$

using the same assumptions as always. The grouping relations of interest are groupings which maximize variation in X_1, X_2, or X_1 and X_2 simultaneously. We have already proven that if, say, X_2 does not belong in the model, then grouping by X_1 is an optimal grouping. But, it is easy to see (once the obvious has been demonstrated) that if X_2 is correlated in the sample with X_1 and has an independent linear effect on Y, grouping by X_1 alone (ranking observations only by X_1 values and grouping adjacent observations) results in a specification error. In fact, with grouping by X_1:

$$E(\bar{b}_1) = \beta_1 + \beta_2 \frac{\Sigma X_1 X_2}{\Sigma X_1^2} \tag{26}$$

and with grouping by X_2:

$$E(\bar{b}_2) = \beta_2 + \beta_1 \frac{\Sigma X_1 X_2}{\Sigma X_2^2} \tag{27}$$

The important conclusion here is the obvious one. In a multivariate model grouping by some concrete criterion which approximates grouping systematically by a subset of the regressors in the micro-model can produce appreciable bias. This is a case of aggregation bias which is directly analagous to and understandable in terms of specification bias. Haitovsky presents an example reproduced in Table 1, using real data where the micro-model takes automobile sales as a linear function of income and automobile inventories. The

aggregation bias which results from the estimation of the relationship from observations grouped by one or the other regressor (rows 2 and 3 of Table 1) are striking.

Table 1 about here

The case we take to be most prevalent -- where the analyst does not have control over the grouping process -- is seriously damaged by the type of error identified by Haitovsky. For cases where the analyst has some control, it is demonstrated that grouping (for our example) from a cross-classification table of X_1 and X_2 results in reasonably good estimates -- no aggregation bias. An example of this approach is shown in the fourth row of Table 1. In addition, if for some reason the analyst had access not to the cross-classification table but to the X_1 table and X_2 table separately, he can solve the pair of equations (26) and (27) for estimates of β_1 and β_2. Haitovsky's example, shown in row five of Table 1 shows that this approach yields good estimates. The reader is referred to the original paper for more detailed justification.

It is best not to concentrate on the technical details at this point since it is enormously important to understand the consequences of the specification error produced by systematically grouping by subsets of micro-regressors. We see immediately that the results of the bivariate case cannot in any way be extended to more useful, more general models. And, we should be alerted that the aggregation consequences in any multi-variate model is likely to be extremely difficult to unravel. We will have more to say on this in the next section.

It appears that the most interesting and informative connection of Theil's auxiliary-regressions with the method of grouping would seem to lie in understanding the specification consequences of any grouping procedure. Although

I have not been able to construct a formal model to support the argument, it seems likely that non-zero coefficients in the auxiliary regressions may be consequences of grouping-specification-error as well as of substantive cross-level effects. We raise this issue because every formulation of aggregation complications except Theil's can very easily (and profitably) be translated into a specification perspective. What we are suggesting here is that additional analysis may support the hypothesis that Theil's result as well depends on grouping which operates selectively on subsets of the micro-regressors.

VI CONCLUSIONS

The results developed in this paper are fairly discouraging from the perspective of a researcher faced with aggregation-disaggregation problems. For, at the present time, there does not seem to be any general methodological solution to the aggregation problem. A number of such solutions have been proposed (see Hannan 1971, Theil 1971) but have turned out not to be very general.[8] Thus the overall picture is a rather gloomy one.

But, we do have more specific knowledge about the mechanisms generating aggregation bias which suggest practical adaptations. It is clear that any successful resolution of aggregation problems in empirical research requires capitalizing on specific features of the model population under study. In other words knowledge about aggregation effects must be supplemented by substantive judgments about likely variation in variables and about causal connections holding among sets of variables. If this is true then there is no general model of how one ought to approach aggregation problems which hold irrespective of the process under study and the concrete features of the empirical research.

It is still important to suggest how specific knowledge of aggregation

effects ought to be employed in a more comprehensive special solution. It is simpler to treat separately the cases where the analyst has control over the grouping relation and the case where the analyst begins with "social structurally" or "naturally" aggregated data.

Random grouping is obviously the safest strategy for the analyst who has control over the grouping procedure. When observations are grouped randomly neither regression nor correlation coefficients are subject to aggregation bias. This is true for multivariate as well as bivariate models. But the analyst pays a price for this assurance. The price is loss of efficiency in estimators. As we noted earlier, sociologists do not typically consider the consequences of such a loss. Yet, as we have shown the loss of consistency (which depends on the "coarseness" of the grouping) can be very sizable with random grouping. In practice this means that the macro-estimates will be unbiased with very large variance (relative to that of the micro-estimator) around the population parameter of interest. Thus even though random grouping avoids bias in the long run (since the expected values of the estimators are equal to the population values of interest), this should not be very comforting in any substantive analysis. That is, the researcher has to live with the estimates he produces with "one shot". Knowing that if he replicated the research an infinite number of times he would make a correct inference in the long run does not provide a great deal of comfort, if the likelihood is great that the present estimate is very far from the population value.

Thus we can specify the consequences of random grouping very precisely. And, as we have seen, grouping by the regressor in a bivariate case or simultaneously by all of the regressors (which is difficult in practice if there are many of them) in a multivariate case biases only correlation coefficients while greatly reducing (relative to random grouping) the loss of efficiency in slope

estimators. As sociologists focus less on correlation coefficients and more on regression parameters, this strategy would seem preferable. But, it is important to point out that this approach is not a "minimax" one. It depends very heavily on the causal assumptions. And, if the knowledge of the causal structure is reliable, this sort of systematic grouping outperforms random grouping. But, if the causal assumptions are wrong, the errors in inferences will be more serious with systematic grouping than with random grouping.[9]

Consider a simple three variable recursive model:

$$x_2 = \beta_{21}x_1 + \epsilon_1$$
$$y = \beta_{y1}x_1 + \beta_{y2}x_2 + \epsilon_2$$

In this model both x_1 and x_2 "belong" in the micro-model and themselves correlated under the specifications of the model. Now consider three types of grouping in this model: (1) random, (2) grouping simultaneously by X_1 and X_2, and (3) grouping by X_1. We know that random grouping while inefficient will not bias either $\bar{b}_{y1}$ or $\bar{b}_{y2}$. Grouping by X_1 and X_2 simultaneously is more efficient and will not bias either $\bar{b}_{y1}$ or $\bar{b}_{y2}$. But, suppose the analyst employs the wrong causal model and ignores the first equation in the model -- that is, assumes that X_1 and X_2 are not linearly related. Under this (wrong) specification he decides to group only by X_1. We have seen that this type of grouping produces an aggregation bias in both $\bar{b}_{y1}$ and $\bar{b}_{y2}$.

This example supports the earlier contention. If the analyst has a good deal of confidence in the substantive assumptions of the model, then the systematic grouping procedure is "optimal". As the level of confidence in these assumptions decreases, random grouping becomes a more and more attractive alternative.

These arguments hold only for <u>unstandardized</u> coefficients. We have seen

the effects of systematic grouping on correlation coefficients and the logic is easily extended to other coefficients standardized to sample variances, e.g. path coefficients. It seems unwise to employ any such standardized coefficients in aggregation situations, but if there is a compelling reason to do so, only random grouping is appropriate.

It has been obvious from the outset that the analysis of already grouped data is a much more difficult undertaking. Here we have argued that the researcher must go through several steps in the analysis. First, the concrete grouping rule must be ascertained. Then, and this is the crucial step, one must make substantive judgments about the abstract variation which is being actualized in the population under study by the concrete grouping rule. To use the substantive example developed earlier, this step involves judging how the mechanism responsible for the placement of children into schools and within schools into classrooms corresponds with variation in other potentially relevant causal variables like parents' social class. Since the number of causal variables which could have been activated is infinite, the search process should proceed by beginning with the variables included in the model. We have seen that the most dangerous (to inference) possibility is that the variation activated by the grouping rule corresponds to variation in the dependent variable. Other possibilities are that grouping might systematically affect variation in single independent (causal) variables in the model, or sets of such variables. Each of these possibilities must be entertained and evaluated on the basis of substantive judgments. They cannot be inferred from the data.

Once we pose the problem in these terms we see additional aggregation bias possibilities. For example, Alker (1969) has developed an argument that grouping might produce joint variation between variables which are not causally related in the micro-model. In such a case, macro-correlations and regressions

can be considered "spurious" from the perspective of inferences to a micro-model. The point again is that when the analyst doesn't control the criteria used in grouping observations, the number and complexity of possible aggregation effects is very troubling. And, we have argued that these types of difficulties will likely be resolved only by detailed substantive investigation of the likely causal connections between grouping variables (as distinct from the concrete grouping criteria) and the substantive variables under study. Since the likelihood of success in such an endeavor seems low, it should probably only be attempted when the only choice facing the analyst is either to make inferences from aggregated data or to abandon the line of inquiry.

One possibly hopeful suggestion from the study of specification bias is that even when the aggregation variables are related to more than one variable in the model, the likely consequences depend on the strength of the causal connections with variables in the model. When observations are "naturally" grouped, we should not expect the grouping to perfectly maximize variation in some other variable. For example, when Blalock (1964) aggregated observations on county level units according to geographic proximity, the aggregation effects were intermediate between the effects for random grouping and systematic grouping by the regressor. We can conceptualize this effect as an imperfect grouping by the regressor or "grouping with error." It is probably safe to assume that all "natural" grouping is to a greater or lesser degree grouping with error. The larger the error component, the closer the grouping is to random grouping (random with respect to the variables under study). But what of the case in, say, a bivariate model where the grouping variable is strongly affecting variation in the regressor and weakly affecting variation in the dependent variable? A generalization of the specification error results would suggest that the degree of aggregation bias would depend, say, for the bivariate case on the ratio of the (unknown) regression coefficients of each of the

substantive variables on the grouping variable(s).

This has not yet been demonstrated formally and is something of a conjecture. Yet, it seems to follow from the known results that if aggregation is simultaneously affecting variation in several variables in a model and the effect on one of the variables is greatly disproportionately larger than on the others, it is reasonable to proceed as if grouping affected only variation in the one greatly affected variable. These sorts of approximate results have not been systematically studied. Since they are very important from the perspective of the practical use of aggregation results, their study should take high priority in further analyses. That is, we need to study (using simulation methodology, analytic strategies and real data analysis) the effects of approximate satisfaction of the conditions for zero or negligible aggregation bias. In other words, we need to know the practical conditions where aggregation bias will be large and damaging to inference.

TABLE 1*

Model	β_1	β_2	R^2
1218 obs.	0.75781 (0.1398)	-0.17778 (0.0367)	.03465
Grouping by X_1	.55051 (1.6139)	.03819 (1.8752)	.72841
Grouping by X_2	-0.65315 (2.5391)	-0.09312 (0.1572)	.90981
Grouping by X_1 and X_2	0.74734 (0.1203)	-0.16242 (0.0323)	.49694
Haitovsky method	0.72713 (0.1033)	-0.17177 (0.0282)	.77045

* Table reproduced from Haitovsky (1966)

FOOTNOTES

*The research reported in this paper was partially supported by National Science Foundation Grant (GS-32065).

[1]This is not to say that the theoretical and methodological issues do not have implications for each other. Quite the opposite. Solutions to the theoretical problems set constraints for approach to analysis problems and the reverse. I have developed this position at some length elsewhere (Hannan, 1971).

[2]This particular complication was suggested to me by Leigh Burstein.

[3]If the information is precise enough to identify individuals, then there is no anonymity. What is needed is characteristics of collectives which will be stable over the period of the study.

[4]These authors address the measurement problem from an aggregation perspective much like that presented in this paper.

[5]The case in which the micro-model is "misspecified" is the most interesting possibility of this type. Grunfeld and Griliches (1960) have discussed the aggregation implications for micro-models which should contain some macro-variables as causal variables. We will discuss specification in Section V.

[6]It is particularly interesting that Theil's most recent treatment of aggregation problems (Theil, 1971) appears in a text in which he also discusses the grouping effects literature. He makes no reference to the other problem in discussing each of these in different chapters.

[7]This is a conjecture which has no formal basis. The conjecture depends on results developed for the aggregation consequences in "random coefficients" regression models (see Zellner, 1969). In this approach one assumes that all micro-units share a common set of response parameters (slopes) but that there is random variability in slopes between micro-units. As long as the variability between micro-units is random, the inter-unit differences cannot be correlated with grouping variables and there should be no aggregation bias. The use of the random coefficients model involves quite restrictive assumptions, however (see Hannan, 1971).

[8]See Theil (1971). The most promising of the suggestions, Theil's "convergent aggregation",uses a random coefficients model as described in footnote 7.

[9]This conclusion is very much like that reached by Blalock, Wells and Carter (1970) in comparing ordinary least squares and instrumental variables estimators in the face of random measurement error. The issue raised is suggested in part by their elegant analysis.

References

Alker, Hayward R., Jr.

1969 "A typology of ecological fallacies." Pp. 69-86 in M. Dogan and S. Rokkan (eds.), Quantitative Ecological Analysis in the Social Sciences. Cambridge, Mass.: MIT Press.

Blalock, Hubert M.

1964 Causal Inferences in Nonexperimental Research. Chapel Hill: University of North Carolina Press.

Blalock, H. M., Caryll S. Wells, and Lewis F. Carter.

1970 "Statistical estimation with random measurement error." Pp. 75-103 in E. Borgotta and G. Bohrenstedt (eds.), Sociological Methodology 1970. San Francisco: Jossey-Bass.

Boot, J. C. G., and G. M. de Wit

1960 "Investment demand: An empirical contribution to the aggregation problem." International Economic Review 1 (January):3-30.

Boudon, Raymond

1963 "Proprietes individuealls et proprietes collectives: Une probleme d'analyse ecologique." Revue Francaise de Sociologie 4(July-September):275-299.

Cramer, J.S.

1964 "Efficient grouping: regression and correlation in Engel curve analysis." Journal of the American Statistical Association 59 (March):233-250.

Duncan, Otis Dudley and Beverly Davis.

1953 "An alternative to ecological correlation." American Sociological Review 18 (December):665-666.

Duncan, Otis Dudley, Ray P. Cuzzort and Beverly D. Duncan

1961 Statistical Geography. Glencoe, Ill.: The Free Press.

Gehkle, C. and R. Biehel

1934 "Certain effects of grouping upon the size of the correlation coefficient in census tract material." Journal of the American Statistical Association Supplement 29:169-170.

Goodman, Leo

1953 "Ecological regression and the behavior of individuals." American Journal of Sociology 18:663-664.

1959 "Some alternatives to ecological correlation." American Journal of Sociology 64(May):610-625.

Green, H. A. John

1964 Aggregation in Economic Analysis. Princeton: Princeton University Press.

Grunfeld, Yehuda and Zvi Griliches

1960 "Is aggregation necessarily bad?" Review of Economics and Statistics 42(February):1-13.

Haitovsky, Yoel

1966 "Unbiased multiple regression coefficients estimated from one-way classification tables when the cross classifications are unknown." Journal of the American Statistical Association 61 (September):720-728.

Hannan, Michael T.

1970 Problems of Aggregation and Disaggregation in Sociological Research. Chapel Hill, N.C.: Institute for Research in the Social Sciences: Methodology Working Paper #4.

1971 Aggregation and Disaggregation in Sociology. Lexington, Mass.: D. C. Heath.

Ijiri, Y.

1971 "Fundamental queries in aggregation theory." Journal of the American Statistical Association 66 (December):766-782.

Johnston, J.

1971 Econometric Methods. Second Edition. New York: McGraw Hill.

Kline, Gerald F., K. Kent and D. Davis

1971 "Problems in the causal analysis of aggregate data with applications to political instability." In J. Gillespie and B. Nesvold (eds.), Macro-Quantitative Analysis. Beverly Hills: Sage Publications.

Leontief, W.W.

1947a "A note on the interrelation of subsets of independent variables of a continuous function with continuous first derivatives." Bulletin of the American Mathematical Society 53:343-350.

1947b "Introduction to a theory of internal structure of functional relationships." Econometrica 15 (October):361-373.

Nataf, Andre

1948 "Sur la possibilite de construction de certains macromodels." Econometrica 16 (July):232-244.

Prais, S. J. and J. Aitchison

1954 "The grouping of observations in regression analysis." Review of the International Statistical Institute 22:1-22.

Robinson, William S.

1950 "Ecological correlations and the behavior of individuals." American Sociological Review 15 (June):351-357.

Shively, Phillips W.

1969 "'Ecological' inference: The use of aggregate data to study individuals." American Political Science Review 63 (December); 1183-1196.

Sono, M.

1961 "The effect of price changes on the demand and supply of separable goods." International Economic Review 2 (September): 239-275.

Stokes, Donald

1969 "Cross-level inference as a game against nature." Pp. 62-83 in Mathematical Applications in Political Science IV. Charlottesville: University Press of Virginia.

Theil, Henri

1954 Linear Aggregation in Economic Relations. Amsterdam: Holland Publishing Company.

1957 "Specification errors and the estimation of economic relationships." Review of the International Statistical Institute 25 (August):41-51.

1959 "The aggregation implications of identifiable structure macro-relations." Econometrica 27 (January):14-29.

1971 Principles of Econometrics. New York: John Wiley.

Yule, Udny G. and Maurice G. Kendall

1950 An Introduction to the Theory of Statistics. London: Charles Griffin.

Zellner, Arnold

1969 "On the aggregation problem: A new approach to a troublesome problem." Pp. 365-374 in K. A. Fox et al (eds.), Economic Models, Estimation and Risk Programming. Berlin: Springer-Verlay.