SOCIAL AREAS: A CASE STUDY IN THE METHODOLOGY OF MASS DATA ANALYSIS*

Dennis McElrath
Department of Sociology and Center for Metropolitan Studies, Northwestern University

Today it is not at all uncommon to find the son of a cattle herder or a peasant seated at a key punch, busily recording population changes in the middle of a government compound in Africa or Asia. As a result, the volume of raw facts on the desk of a social scientist has increased enormously in the last decade. But this quantitative growth in information is the product not only of advances in data-accumulation and processing techniques; it also stems from the increased bureaucratization of the world. Not just technological innovations, but changes in the organization of society as well sponsor the diffusion of social bookkeeping.

The enlarged scale of government and of business enterprise has increased the demand for detailed information. Much of this information is now gathered on a regular basis: censuses, vital statistics and labor force data are prepared every decade, or yearly, or quarterly, or monthly. Thus, (1) a greater volume of mass data now is accumulated from all over the world (2) on a greater variety of topics (3) at regular intervals. The advantage of mass data analysis, for both the social scientist and the policy maker, lie in these three characteristics of mass data: volume, variety and regularity.

Social bookkeeping records are not ordinarily maintained to answer specific policy issues. The mass data analyst, when facing a particular question, must therefore select relevant data out of a plethora of periodic observations and somehow reduce them to manageable proportions. Currently three major approaches are used to handle this problem of selection and data reduction. These procedures are: correlation analysis (such as multiple regression analysis and multiple factor analysis); classification; and construction of typologies.

The correlation approach frequently is used here where (1) little is known about the phenomena under investigation, or (2) where a strong inductive tradition exists in a discipline, such as in engineering or medicine. Correlation techniques attempt to reduce a variety of information to a few critical variables by observing the way in which a large number of variables are interrelated. By examining their inter-correlations, it is possible, with present computer techniques, to reduce these many measures to a few sets of variables which are closely related to one another. These few combinations of sets are then analyzed, treating each set as if it were a single variable. A mass of information is thus reduced to manageable dimensions.

The relevance of a set of variables for the analysis of a specific problem is determined by following similar procedures. Strategic variables are identified through a correlation analysis which determines how well a particular measure or combination of variables predicts any specified "dependent" variable. Thus, correlation is used both to reduce information and to determine its relevance.

The problems which plague this "shotgun" approach, of course, are that (1) relevance is established by correlation but correlation is not a cause and does not necessarily lead to understanding; and (2) that there is no way of determining whether or not reduction by correlation has obscured the critical variables by mixing them with related but unimportant measures. Thus, while data are reduced and predictions calculated understanding is not always advanced by this approach.

Development of classification schemes often attempts to instill explanation and understanding into mass data reduction operations. Here, a large number of observations are reduced by selecting from and combining original categories guided by some frame of reference. The approach is deductive in that the analyst picks a few observations or sets of observations out of many because his theory or explanatory frame of reference suggests that these are critical. His bent is toward explanation and to variables with strong interpretive power rather than to the empirical regularities guiding correlation analysis. Thus, a thick volume of occupational titles may be re-classified to a simple dichotomous classification of manual and non-manual occupations based on a theory indicating that this split in the skill hierarchy is critical in the development of nations.

Construction of typologies usually follows the same theoretical bent as classification. However, reduction of mass data to typologies has one advantage over classification in that data are reduced to several independent dimensions whose joint or combined variation is used to interpret a policy question. Because of this they are suited to the more complex theories now current in the social sciences. Their construction, however, is less standardized than the fairly automatic numerical reductions by correlation or factor analysis. All three are designed to handle a large volume and great variety of data. Differences between them lie in their theoretical relevance and in the extent to which a catalog of appropriate techniques is available.

The volume, variety and regularity of mass data have their disadvantages though they do offer opportunities to social scientists and to policy makers. The analyst usually must accept as given the observations, categories and areal units used in social bookkeeping records. These data are highly structured before they reach the analyst and this pre-structuring strongly influences the kinds of understandings or recommendations he can make. First, the observations themselves suffer from a variety of levels of interpretation which occur throughout the data accumulation process. These range from obvious instances of interviewer bias to subtle interpretations used in fitting a perceived world into prearranged categories. Second, often mass data are classified in categories derived from bookkeeping traditions of a data-gathering institu-

tion. Continuity or comparability of categories frequently is more important than current relevance. In addition, classifications may be designed to serve diverse interests of various agencies, bureaus or nations and relevance to a particular issue at times may be obscured for the sake of generality. The development of standard classifications and classification procedures which are widely used throughout the world has yielded fairly reliable information of great generality, but sometimes at the cost of some fairly revealing local classification schemes.

Legal restrictions also often constrain data presentation. Most censuses must guarantee the anonymity of respondents in their published records so that categories must be fairly large and cross tabulations limited. Otherwise an individual could be singled out from the crowd.

The third problem with mass data arises from the fact that they are collected on the basis of areal units and are tabulated areally. Nearly all censuses collect observations using a small local area as a basis for assigning individuals or households to an enumerator. Data usually are summarized for this enumeration area or for a set of such territories. If the areal unit is large enough, a great deal of material may be summarized and cross-classified while the anonymity of individual residents and households is preserved. Thus a great deal may be learned from census publications about a particular census tract, but nothing may be learned about a particular individual.

Some students of the city have attempted to talk about individuals from correlations of areal data. They have reasoned (erroneously) that these ecological correlations are closely analogous to individual correlations. On the other hand, there have been serious attempts to develop forms of analysis where the local area is the basic unit of observation and interpretation. This sociology of locality groups attempts to avoid misuse of ecological correlations and still take advantage of the wealth of local area data presented in almost all large scale social bookkeeping accounts.

Today I should like to report on several interesting recent developments in social area analysis.[1] These modifications and changes which I shall describe stem directly from attempts to take advantage of the increased volume, variety and regularity of local area data now available in national censuses of metropolitan areas throughout the world. The history of these changes represents a case study of more than a decade of effort on the part of a number of scholars to develop an approach to mass data analysis which takes advantage of these prevalent characteristics and also provides a theoretically meaningful and empirically grounded frame of reference for interpreting these data. Today I can only highlight some of these developments.

Most of you are familiar with social area analysis in the form originally applied to Los Angeles in 1949 and elaborated by Shevky and Bell in 1955.[2] In this second monograph, which included a number of revisions of the earlier work, the authors rigorously describe how census tract populations may be located in a "social space" or three-dimensional typology. This space is defined by three axes along which resources and opportunities are distributed in modern society. The location of tract populations along each axis is determined by combining several standardized ratios computed for each tract. "Urbanization", the first dimension of the typology arrays subareas according to prevalent alternative styles of family life ranging from "familism" on the one hand to "urbanism" on the other. It is measured by subarea distributions of fertility, women at work and house type. Thus, a familial area is characterized by a low proportion of women at work, a high fertility ratio and a high proportion of single family dwellings. "Social rank", the second dimension, arrays subpopulations by distributions of literate and non-manual skills. It is measured by combining education and occupation ratios. Finally, "Segregation" arrays subareas by using tract measures of the distribution of racial and nationality groups living in relative isolation.

From its inception, then, social area analysis (1) uses a typology to compress a large variety of widely recorded population characteristics; (2) forms this typology by selecting from and combining these characteristics guided by a general theory of social differentiation which may be broadly applied; (3) views local area populations as fundamental units of observation, and (4) interprets variations in local area populations in terms of combined importance of their location along all three axes of social differentiation. These assets of the original formulation make the approach highly generalizable and suggest that its development may have broad implications for analyses of mass data in a variety of quite different settings. Subsequent modifications represent then, an important case study in the methodology of mass data analysis.

A survey of revisions and modifications in social area analysis reveals that significant changes have occurred in five major areas. First, there have been important changes in the dimensions themselves. These occurred in response to observed empirical regularities and, more importantly, in several attempts to achieve greater theoretical clarification than existed in the original formulation. The segregation dimension is the major point where both of these considerations are operative.

In the original Shevky formulation, tract populations were arrayed according to the distribution of "subordinate" groups. The general view was that a large scale society included within its social boundaries a variety of populations of varying ethnic, racial and national backgrounds. Subordination was the outcome of their differential date of entry into the social system as well as shared physical and cultural visibility. Spatial isolation is viewed as a conse - quence and perpetuator of subordination. An em-

pirical observation occurred in a social area analysis of Accra, Ghana, which stimulated further exploration of this dimension.[3] It became clear that here segregation involved two quite different conceptual dimensions which are usually compounded in the American situation.

Urban subpopulations in Accra were differentiated by tribal origin, on the one hand, and by migration experience on the other. While these frequently are empirically closely intertwined, they would seem to have quite different theoretical implications -- for example, assimilation of migrants involves socialization of the urban scene while ethnic assimilation involves the erasure of social stigmata. Yet both of these differentiators are involved in the single dimension of segregation. These empirical observations coupled with a series of comparisons conducted by John Barkey relating segregation to the assimilation process in Chicago led to an exploration of the theoretical import and analytic utility of separating ethnic status -- based on physical and cultural visibilities -- from migration status -- based on such shared migration experiences as the volume, variety and intensity of migration, the steepness of social boundaries crossed in the process of migration; and structural differentials in time of arrival.[4] By treating migration status separately from ethnic status, it is possible to examine situations where the statuses are compounded (e.g., recent Negro arrivals in Chicago), as well as their separate occurrence among subpopulations (e.g., long term resident Negroes in Chicago or recent rural-to-urban migrants in Mexico City). Thus empirical observations stimulated examination of these conceptually distinct dimensions. This separation modifies somewhat the underlying theory of social differentiation and suggests that the original three-dimensional typology be supplanted by dividing segregation into ethnic status and migration status. Work along these lines is continuing.[5]

A second modification of social area analysis concerns index construction. A basic requirement of the indexes used to measure each dimension of the typology is that they range evenly across the variety of situations studied. This means that each index should be "univocal" (i.e., measure the same presumed reality in each test situation) and that the value of the index (index score) be comparable across all urban areas. When the approach was applied across a variety of situations, several modifications of the indexes were required. One such modification occurs when a component is inoperative in a particular situation. The application of the urbanization dimension to Rome is an example of this.[6] One of the three components of this axis is house type. In the United States, the proportion of all housing units which are single family units is used as a partial indicator of life style. Obviously this index is not useful where the opportunity for subpopulations to be distributed in single units is non-existant. In this instance, between tract variation in house type is negligible. In fact, the slight variation which is observed is probably not indicative of life style at all but rather of the distribution of a few shacks (barrachi) occupied by poor migrants. That is, it probably measures migration status and social rank rather than urbanization. In this situation all that can be done is to throw out this component and perhaps substitute some other measure of this aspect of urbanization. I, personally, doubt that this measure is essential to the underlying concept of urbanization and probably could be discarded in all situations. Some of the recent work of Theodore Andersen suggests that this may indeed be the case.

Somewhat similar index problems occur in applying the education ratio (a component of social rank) in cross-societal studies. However, in this instance, functional literacy seems to be the operative concept and may be handily measured from most census sources.[7]

There is an additional index problem which may be partly involved in the above difficulties. This concerns the extent to which any index of social structure may be "culture free". This is, of course, a question which should be examined empirically whenever an index is applied across societies. The relevant proposition guiding such research using the Shevky frame of reference is that all large scale urban societies share fundamental social structures and that these in turn are subject to common measurement. A partial substantiation of this position is observed when all components of each axis vary together in an expected pattern across a set of subareas. When they do not, one should examine both the scale of the society as well as a cultural contamination of indexes. It may well be that in a small scale society, a particular form of differentiation is inoperative. In this situation, it may also occur that a particular index is inoperative or operates at variance with expectations because of the culture-bound social meaning of the index component. This latter difficulty is the case with house type in Rome, while the former occurs with the use of women in the labor force in Ghana.[8]

A third way in which the Shevky approach has been modified involves methods of standardizing components of each axis. Since Professor Orleans is going to discuss this subject in some detail, I shall merely mention that the question of standardization has not been resolved and we have therefore built a good deal of flexibility in this area into our computer programs.

Dr. Orleans will also discuss several solutions to problems encountered in shifting from one areal unit of analysis to another -- the problem of fit, for example. This is a fourth way in which the social area approach has been modified. I should like to preface his thoughtful paper with the single observation that, unlike the traditional "natural area" concept of classical ecology, social area analysis does not assume homogenous local area populations. Rather, the assumption is that each local area, census tract, enumeration area, gruppi di sezioni, precinct or what have you, is characterized by a distribution of attri-

butes which may or may not vary systematically within a local area. It is this distribution which is either the object of interpretation or defines the context within which other dependent variables are interpreted.

Finally, a major area of change in social area analysis has been the development of a tool kit of supplementary techniques which facilitate application of this approach. I shall merely list them.

CENSAN - a flexible program adapted to the 709 which yields social area distributions from a variety of local area data. Options for standardization of components and the segregation dimension are built into this program.[9] Figure 1 presents sample output of this program based on tract statistics from ten metropolitan areas.

RATIO - a flexible program which yields a variety of local area ratios which may be used in conjuntion with CENSAN.

SYMAP - a computer graphic technique developed by Professor Howard Fisher which supplements CENSAN and yields highly legible social area topographies indicating the relative intensity of local areas within the social area grid. Figure 2 presents sample output from this program.

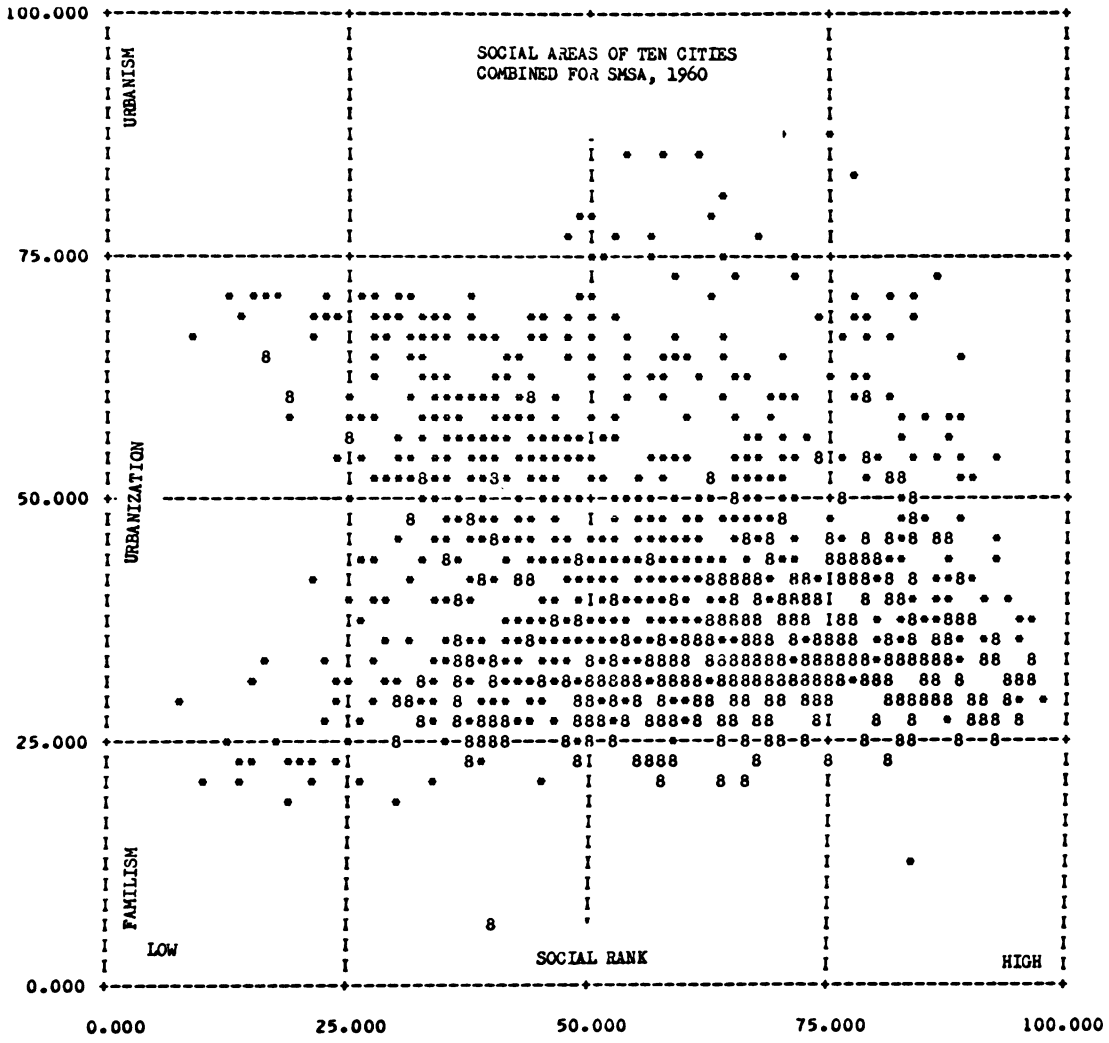
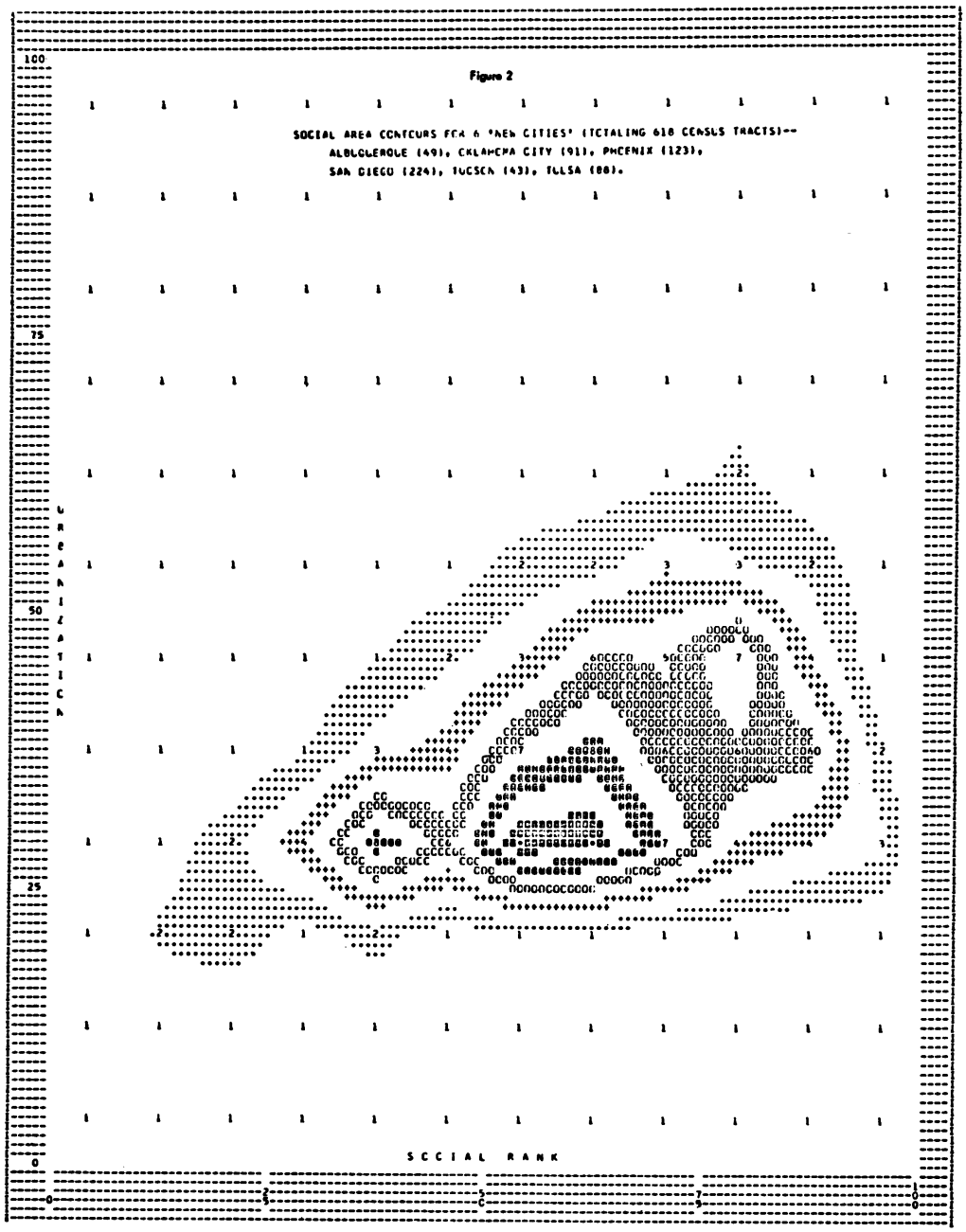
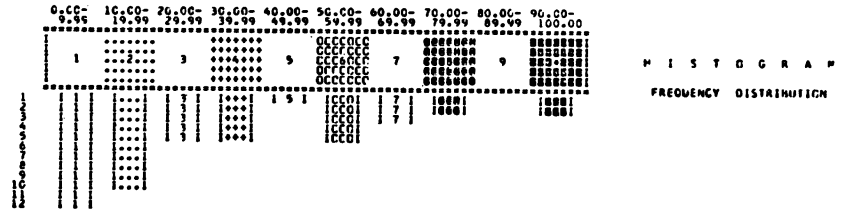


Figure 1



BASED ON STANDARD METROPOLITAN STATISTICAL AREAS, 1960 CENSUS.
SCALE IS IN PERCENTAGE DECILES— ABSOLUTE RANGE 0 TO 41.
'SOCIAL RANK' = A STANDARDIZED INDEX BASED ON EDUCATION AND OCCUPATION.
'URBANIZATION' = A STANDARDIZED INDEX BASED ON EMPLOYMENT OF WOMEN, MULTIFAMILY HOUSING, AND INFERTILITY.
(HOWARD F. FISHER)



In addition to several special programs, several special programs, several non-computerized techniques have been developed as well. These include a technique for sampling local areas by distance and direction weighted for the typical density gradient of single centered metropolitan areas. This facilitates initial exploration of relations between geographical distributions of local areas and their configurations in social space. In all of these efforts we have tried to develop tools which are sufficiently flexible to take advantage of the increased volume and variety of local area data and which will be ready for the 1970 round of censuses. I should greatly appreciate hearing your comments on these revisions before that date.

* Paper prepared for delivery at the annual meeting of the American Statistical Association in Chicago, December 29, 1964. The first section was prepared with Raymond W. Mack.

1. An annotated bibliography of work using this approach is available from the author.
2. Eshref Shevky and Marilyn Williams, The Social Areas of Los Angeles: Analysis and Typology, Berkeley and Los Angeles: University of California Press, 1949; and Eshref Shevky and Wendell Bell, Social Area Analysis, Stanford University Press, 1955.
3. Dennis C. McElrath, "Social Change and Urban Social Differentiation: Accra, Ghana", (Mimeographed).
4. John W. Barkey, "Selected Aspects of Change in the Social Areas of Chicago, 1930-1960: A Research Proposal". Unpublished Master's Thesis, Department of Sociology, Northwestern University, Evanston, Illinois, 1963.
5. Dennis C. McElrath, "Migration Status in Accra, Ghana", (Mimeographed).
6. Dennis C. McElrath, "Social Areas of Rome: A Comparative Analysis", American Sociological Review, Vol. 27, No. 3. June, 1962, pp. 376-391.
7. Loc. cit.
8. Dennis C. McElrath, op. cit., "Social Change and Urban Differentiation: Accra, Ghana".
9. Program available on request from the author.