THE CONSTRUCTION OF "SOCIAL AREA" INDICES: AN EVALUATION OF PROCEDURES[1]

By: James M. Beshers - Purdue University

Shevky, Bell and Tryon have described several procedures for constructing "social areas" from census tract data.[2] These procedures all consist of reducing the demographic characteristics of census tracts to a relatively small number of "dimensions." Various statistical techniques may be employed in this reduction - factor analysis, cluster analysis, Guttman scaling, etc.

Evaluation of these procedures has been undertaken by several authors recently.[3] Most of these investigations consist of replications of "social area" procedures on a number of cities. Evaluation of the procedures is therefore based on the empirical results of using them. A consistency, or reliability, criterion for the procedures is implicit in this work. But one cannot determine from such studies whether lack of consistency reflects deficiencies in the procedures themselves, or variation in social characteristics among U.S. cities.

An alternative approach is to examine the procedures themselves. The mathematical properties of these procedures are all well known - a comparison on this basis is unnecessary. However, one may examine critically the use of the procedures in this particular empirical problem. To do so, one must follow through the details of applying these procedures to at least one empirical case.

The present study is directed to this alternative approach using data from Cleveland, Ohio, in 1950.[4] The results reported here are part of a larger study of statistical procedures used in manipulating census tract data. The choice of Cleveland and the other constraints imposed by the larger study design do not appear to the author to have distorted the results of this part of the research.

STUDY DESIGN

In this study twenty variables were selected from the census tract data. The relationships between these variables were used as a criterion for reducing the twenty variables to a smaller number of constructs - thereby eliminating redundancy. Subsequent calculations are simplified by obtaining a small number of constructs. Further, the theoretical significance of the raw data is summarized in the constructs; thus the subsequent analysis will have direct theoretical relevance. Indices for these constructs will be chosen.

On the basis of exploratory calculations, eighteen tracts were eliminated from the analysis of the 1950 data. Fifteen tracts had less than six hundred inhabitants, while three tracts had large institutional populations.

An intercorrelation matrix was computed for all twenty variables over the 331 remaining tracts. Two techniques were used to analyze the correlation matrix. An effort to reveal the effects of a single factor was made with Guttman s simplex approach.[5] This approach consists of rearranging the correlation matrix so that the largest correlations are all next to the main diagonal; the correlations should decrease in the successive diagonals away from the main diagonal and achieve the smallest value in the extreme corner of the matrix. Other methods of inspecting the correlation matrix were used to supplement the rearrangement of the columns. Table 2 contains the analysed matrix. In order to clarify the relationships a number of the variables have been reversed. For example, X_3, percent non-white, becomes percent white in order to emphasize the positive relationship with the socio-economic status indicators.

The correlation matrix was also analysed by factor analysis. The principal components technique was used. Communalities were estimated by calculating the inverse matrix and using the diagonal elements to obtain the square of the multiple correlation coefficient of a particular variable on the other nineteen variables. Guttman has shown that the communality is an upper bound for the multiple correlation.[6] Therefore the multiple correlation, as a lower bound for the communality, may be used as an estimate of the communality. The rotated factors are reported in Table 3.

MATRIX ANALYSIS

A detailed examination of Tables 2 and 3 will reveal a surprising correspondence between the results of the two methods of analysis. The arrangement of Table 2, following the simplex rule, brings out the dominance of socio-economic status. The variables exhibiting the strongest effect of this dimension are at the left, arranged in approximate order of decreasing effect. If only the first ten variables are examined as a group, an approximate simplex can be seen. With the exception of X_{17}, the large correlations are near the main diagonal, and a general decline in correlations can be seen away from the main diagonal. The next four variables in Table 2 have a general positive relation with the social status group, but no simple pattern is revealed by the magnitude of their coefficients.

The last six variables have little or no consistent relationship with the major socio-economic variables. Four of these, however, have large positive intercorrelations. They seem to represent a life cycle stage in residential areas.

These two main effects in Table 2 have been headed A and B. A further phenomenon of interest can be uncovered by examining the correlations for X_3, the race index. There seems to be a moderate average correlation with the social status group, about .4, but there are high correlations with X_{14} and X_{12} among the social status indicators. That is, non-white is correlated

TABLE 1

List of Variables, Designations when Reversed, and Their Derivation from Raw Data

X_1--Males (females)--Percent males--male/total population

X_2--foreign born white (native born white)--percent foreign born white--foreign born white/total white

X_3--race (white)--percent nonwhite--nonwhite/total population

X_4--large family (small family)--population per household

X_5--education--median school years completed

X_6--same house--percent same house in 1949--same house in 1949 as 1950/persons one year old and over, 1950

X_7--income--median family income

X_8--children--percent males under 15 years--males under 15 years/total males

X_9--old people (not old)--percent males over 65 years--males over 64 years/total males

X_{10}--males married--percent males married--married males/males 14 years old and over

X_{11}--females working (females not working)--percent females in labor force--females in labor force/females 14 years old and over

X_{12}--unemployment (employment--percent males unemployed--unemployed males/males 14 years old and over

X_{13}--professional--percent males professional--male professionals/employed males

X_{14}--laborers (not laborers)--percent males laborers/employed males

X_{15}--owner occupied--percent owner occupied dwelling units--owner occupied dwelling units/all dwelling units

X_{16}--one dwelling unit--percent separate dwelling units--one dwelling unit, detached/all dwelling units

X_{17}--old dwelling units (new dwelling units)--percent dwelling units built 1919 or earlier--dwelling units built 1919 or earlier/number reporting age

X_{18}--crowding (not crowded)--percent 1.01 or more persons per room--number dwelling units 1.01 or more persons per room/number reporting persons per room

X_{19}--median rent--median contract monthly rent with estimates from X_{20}

X_{20}--median value of dwelling units--median value of owner occupied one-dwelling-unit structures with estimates from X_{19}

TABLE 3

First Four Factors: Rotated and Interpreted

	Social Status	Young Family	Females Stability	Race
X_5	.65	.23	-.07	.31
X_{13}	.79	.31	.01	.38
X_{20}	.88	.16	-.01	.29
X_{19}	.77	.22	-.07	.42
X_7	.85	.03	.12	.01
X_{14}	.79	.40	-.04	-.28
X_{12}	.83	.04	.12	-.28
X_{17}	.79	-.29	-.17	.10
X_{16}	.78	-.34	-.25	-.14
X_{15}	.83	-.29	-.04	-.26
X_{18}	.66	.47	.20	-.26
X_{10}	.72	-.39	.21	.05
X_2	.61	.04	-.29	.04
X_3	.58	.38	-.00	-.61
X_1	.32	-.12	.66	.29
X_{11}	.54	-.62	-.05	-.09
X_8	.22	-.85	-.07	-.04
X_9	.13	-.70	-.19	.02
X_4	.01	.87	-.06	-.05
X_6	.09	-.26	.78	-.30

TABLE 2

Intercorrelation Matrix of Twenty Tract Variables with Derived Constructs

	A: General Social Status										C					B				
	A_1					A_2		A_3												
	Social Status					$A \neq C$		$A \neq B$			$A_2 \neq C$	$A \neq B$	A	Race	Only $X_{10} \neq X_6$	Young Family				Only $X_1 \neq X_4$
	X_5	X_{13}	X_{20}	X_{19}	X_7	X_{14}	X_{12}	X_{17}	X_{16}	X_{15}	X_{18}	X_{10}	X_2	X_3	X_1	X_{11}	X_8	X_9	X_4	X_6
	Ed	Prof	DU Val	Rent	Inc	Not Lab	Emp	New DU	One DU	Ownr Occ	Not Crd	M Mar	Nat Wh	Wh	Fem	F Not Wrkg	Chil-dren	Not Old	Large Fam	Same House
X_5		.89	.84	.82	.70	.70	.63	.62	.53	.53	.58	.54	.64	.42	.32	.29	.04	-.02	-.21	-.15
X_{13}			.88	.84	.70	.60	.54	.55	.47	.46	.57	.40	.52	.36	.31	.25	-.08	-.18	-.22	-.09
X_{20}				.86	.79	.63	.65	.71	.63	.65	.59	.53	.49	.37	.29	.37	-.00	-.04	-.08	-.03
X_{19}					.71	.56	.52	.65	.48	.48	.48	.46	.46	.27	.26	.26	-.05	-.05	-.18	-.17
X_7						.66	.67	.65	.64	.72	.61	.59	.42	.49	.28	.46	.12	.03	.02	.19
X_{14}							.77	.48	.45	.57	.80	.47	.52	.81	.18	.18	-.11	-.10	-.39	-.08
X_{12}								.57	.62	.73	.69	.63	.50	.67	.29	.43	.15	.10	-.07	.23
X_{17}									.72	.74	.34	.67	.45	.28	.16	.61	.39	.37	.23	.00
X_{16}										.91	.34	.57	.51	.35	.07	.65	.40	.30	.32	.04
X_{15}											.50	.63	.41	.46	.16	.61	.32	.27	.30	.22
X_{18}												.35	.29	.68	.19	.06	-.30	-.26	-.40	.16
X_{10}													.42	.23	.54	.61	.51	.46	.24	.23
X_2														.39	.12	.31	.23	.14	-.17	-.22
X_3															-.04	.16	-.08	-.22	-.41	.16
X_1																.13	.16	.07	.09	.45
X_{11}																	.70	.40	.55	.23
X_8																		.68	.71	.19
X_9																			.54	-.04
X_4																				.28
X_6																				

The effect of a general status factor can be seen by the magnitude of correlations reading from left to right, a modified Guttman Simplex. However, note the composites, A_2 and C link together X_{14}, X_{12}, X_{18}, and X_3; while A_3 and B link together X_{17}, X_{16}, X_{15}, X_{10}, X_{11}, X_8, X_9, and X_4.

with laborers and unemployment more than might be expected from a general social status effect. This result is in keeping with the findings of sociological research with individuals. Further, the high correlation with X_{18} indicates that non-white is associated with crowding. The designation C has been given to race in the headings of Table 2.

A linkage between A and B may be discovered by scanning the correlations in B. While most of the correlations of B with A are negative, the average correlation of B with X_{17}, X_{16}, and X_{15} is above .3. This linkage of life cycle stage with characteristics of dwelling units is not strong, but it is significant information with regard to the nature of the young family group in the matrix. A similar linkage can be seen between B and X_{10}.

Only two variables, X_1 and X_6, have little to do with A, B, and C. In fact, they have almost as high a correlation with each other as they do with any other variable. Further, there is no clear theoretical reason for the relationship between these two variables, suggesting that they may be excluded from subsequent consideration.

The combined simplex and subsequent analysis suggests that three constructs may represent the relationships in the correlation matrix - socio-economic status, young family life cycle stage, and race. The first and last of these three are familiar constructs in social structure analysis. The life cycle stage is not readily assimilated to social structure, although it may be seen as a status associated with the nuclear family. The life cycle stage may be regarded as a construct for urban social structure analysis, but further research and development of theory are necessary if its relevance to social structure in general is to be discovered. One reason for the neglect of this construct is suggested by the correlation matrix. Except for the relationships discussed above, B is almost independent of both A and C.

The same constructs are evident in the factor analysis. Five factors, representing 97.4 percent of the total estimated communality, were considered for rotation. The first two factors, representing 53.6 percent and 22.1 percent of the total estimated communality, were identifiable without rotation. Reference to Table 3 indicates that these are constructs A and B from the previous analysis. The positive value of X_4 on the second factor is due to the fact that it was reversed in this analysis. The third and fourth factors, representing 9.3 percent and 7.6 percent of the total estimated communality, were clarified by an orthogonal 45 degree rotation with reference to each other. The third factor consists of the relationship between X_1 and X_6. The fourth factor represents race predominantly, with scattered relationships with the other variables. The fifth factor, representing 4.8 percent of the total estimated communality, had low factor loadings on all variables, and showed little possibility of clarification by rotation. It was dropped from the analysis.

COMPARISON WITH OTHER STUDIES

The results of this construct formation process for Cleveland are quite similar to results obtained by Shevky and Bell for Los Angeles, and Tryon for San Francisco.[7] However, there are important differences in detail. The Shevky and Bell constructs of economic status and family status are very similar to the socio-economic group and the young family group isolated in the analysis above. However, the mode of presentation and the method of analysing the relationships between the Shevky and Bell constructs obscures the relationships between the constructs. The linkages reported between A and B in Table 2 are evidently unknown to these investigators. Bell uses factor analysis with a greatly reduced matrix of variables so that these results cannot be seen clearly.[8]

The Shevky and Bell ethnic status construct includes nationality as well as non-white as a component. The analysis for Cleveland shows that foreign-born white and non-white are about equally related to socio-economic status, but the correlation between the two is only .39, suggesting that they should not be combined. Again, the linkages between A and C are not reported by Bell.

Tryon's constructs differ largely from Bell's in that economic status has been replaced by socio-economic independence and socio-economic achievement. Since the correlation between these two constructs is .9, one may question the usefulness of separating them. Otherwise the socio-economic status and young family group found in the present analysis are equivalent to Tryon's constructs of combined socio-economic status indicators and family life. Tryon's assimilation construct includes race along with other related variables, and therefore is roughly comparable to the single item of race.

The linkage between race and laborers in Table 2 may be seen in the components of assimilation. Other linkages from Table 2 may also be found by cluster analysis; correlations may be found between clusters, e.g., family life is correlated .3 with assimilation. But in two respects the method of cluster analysis obscures the relationships between the variables. First, cluster analysis makes no provision for reordering the matrix to demonstrate the effect of one general factor. Because of this lack Tryon's correlation matrix cannot be scanned intelligently; Cluster I, Cluster V, and Cluster VII are all made up of socio-economic items and are very similar, yet they are scattered from one end to the other of the correlation matrix, as their numerical titles suggest.[9] Second, the only method of investigating linkages is the correlation between clusters; this method will overlook cases in which a single variable links two clusters together. Perhaps cluster analysis, by relying on average correlations to rearrange a correlation matrix, has become too rigid a set of rules. Tryon's inability to incorporate other simple techniques in organizing and analysing a correlation matrix is responsible for the neglect of linkage analysis.

INDEX CONSTRUCTION

The choice of indices for the three constructs - socio-economic status (or "social status"), young families, and race - was guided in the present study by two criteria. The statistical relationships in Table 2 and Table 3 were used to determine the importance of variables within the construct groups. Concurrently, the tabulations of the 1940 and 1930 census were used to discover comparable indices in the earlier data. Therefore both a statistical criterion and a comparability criterion were employed.

The approach to index construction used here differs from the usual psychological testing approach. Several variables could be readily weighted and combined to form multi-variate indices for each construct. Instead attention was focussed upon selecting a single variable which could best meet statistical criteria for representing a construct yet could also meet other criteria as well. This procedure has several desirable characteristics. First, comparability with previous census data is more easily achieved. Second, comparison with other studies using 1950 census data is direct; one may test these indices for other cities, select those cities for which the same variables are valid and then compare these cities in terms of the census variables rather than a specially contrived index. Third, in case of complex findings the simpler index should facilitate interpretation. In general, arbitrary operations upon the raw data are likely to obscure subsequent analysis and interpretation. The Shevky and Bell choices of particular cities and censuses to define their Social Area Indices seem to possess this arbitrary quality.[12]

The dwelling unit value with estimates from rent, X_{20}, had the highest intercorrelations among the first five indicators of socio-economic status, and had the largest factor loading on the social status factor. In the 1930 tract data for Cleveland, equivalent monthly rental figures computed by Howard Whipple Green are the only data which correspond to the socio-economic data published in 1950.[10] From these two points of view it seemed that X_{20} should be used as an index of socio-economic status for the 1950 analysis of spatial distribution. A complication arises from this choice. Hoyt's theories and research on urban areas utilize rent as a general socio-economic index.[11] Therefore tests of Hoyt's work with 1950 data should be guided by the rent variable, X_{19}. Since the two variables were constructed as composites, and there is a corelation of .86 between them, little distortion will result from this procedure.

In the young family group children, X_8, has the highest average correlation with the other variables. Age distributions are reported in the 1930 and 1940 censuses, and the classification permits the construction of the same variable as X_8, was selected as the index of the young family group for subsequent analysis.

The index of race, X_3, is itself a single variable, and is tabulated in the 1930 and 1940 censuses. The adoption of this variable as an index for subsequent analysis leads to complications. Percent non-white, X_3, is extremely skewed. There is little to be gained by transforming this variable since fifty-nine percent of the tracts have the value zero. Therefore it may be used in subsequent analysis in its present form. Several techniques of analysis assume the use of normal variables; indeed, the inclusion of this variable in the correlation matrix carried the assumption of normality with it.

SUMMARY

The present critical evaluation may be summarized as follows. The work of Shevky and Bell and of Tryon has led to the identification of three constructs for summarizing the characteristics of census tracts. One of their constructs is novel to the literature and poses a number of interpretation problems. The other two, while familiar, are nevertheless dramatized by their analysis. However, the techniques used in identifying these constructs have overlooked a number of important points of detail, especially some of the more intricate relationships between the constructs. Simplex methodology and systematic inspection of the correlation matrix lead to these conclusions.[13] Finally, the choice of indices for these constructs and some alternative indices are suggested. More empirical cases would be needed to determine the substantive weight of these criticism. Yet even without further research important procedural defects in Social Area analysis are indicated here.

FOOTNOTES

1. This research is drawn from the author's unpublished Ph.D. thesis, "Census Tract Data and Social Structure: A Methodological Analysis", University of North Carolina, 1957. Daniel O. Price, Rupert B. Vance, and James A. Norton, Jr. have helped to formulate these ideas. A grant from the Purdue Research Foundation provided writing time.

2. Eshref Shevky and Wendell Bell, Social Area Analysis, (Stanford: University Press, 1955), Robert C. Tryon, Identification of Social Areas by Analysis, (Berkeley and Los Angeles: University of California Press, 1955).

3. Maurice D. Van Arsdol, Jr., Santo F. Camilleri, and Calvin F. Schmid, "The Generality of Urban Social Area Indexes", American Sociological Review 23 (June 1958).

4. U.S. Bureau of the Census, U.S. Census of Population: 1950, Vol. III, Census Tract Statistics, Chapter 12. Washington: U.S. Government Printing Office, 1952.

5. Louis Guttman, "A New Approach to Factor Analysis: The Radex", Chapter 6 in Paul F. Lazarsfeld (ed.) Mathematical Thinking in the Social Sciences, (Glencoe, Ill.: The Free Press, 1954).

6. Louis Guttman, "Multiple Rectilinear Prediction and Resolution into Components", Psychometrika, 5 (1940), pp. 75-100.

7. Shevky and Bell, op. cit., p. 4, Tryon, op. cit., p. 16.

8. Wendell Bell, "Economic, Family, and Ethnic Status", American Sociological Review, 20 (Feb., 1955).

9. Tryon, op. cit., p. 54 (insert).

10. Howard Whipple Green, Population Characteristics by Census Tracts, Cleveland, Ohio, 1930, (Cleveland: Plain Dealer Publishing Co., 1931).

11. Homer Hoyt, The Structure and Growth of Residential Neighborhoods in American Cities.

12. Shevky and Bell, op. cit., pp. 67-68.

13. A similar methodological statement is contained in Edgar F. Borgatta, "On Analyzing Correlation Matrices: Some New Emphases", Public Opinion Quarterly, XXII no. 4, (Winter, 1958-59).