CLUSTER ANALYSIS: AN APPLICATION TO TYPOLOGY OF URBAN NEIGHBORHOODS

Herbert Bixhorn, District of Columbia Government

1. Introduction

Cluster analysis is the name given to a body of methods for partitioning a heterogeneous collection of objects into groups or clusters in which the objects tend to be similar. In this paper a particular type of cluster analysis is introduced and applied to the problem of classifying geographic subareas of a city into a meaningful typology. The objects to be classified here are census tracts of a city, each tract having a set of variables associated with it. Tracts are considered to be similar or to belong to the same cluster if their values on these variables are similar according to some criterion. The description of the criterion function used in clustering will be more conceptual than rigidly mathematical. The reader who is acquainted with matrix algebra will find a complete discussion of the subject in Rubin and Friedman [1]. The main purpose of this paper is to show some of the advantages of one type of cluster analysis over methods now in common use. To aid in this, both artificial examples and results of analyses performed on tracts of Washington, D. C. will be given.

Before discussing the method of clustering used in this paper, we will review two commonly used methods of classification: summed-ranks and principal components. This will give some indication of the problems encountered in classification.

2. Summed-ranks

The method of summed-ranks will be introduced by first discussing the method of ranking on one variable.

EXAMPLE: Suppose we wish to partition a set of 12 census tracts on the basis of median family income. (See Table 2a at end of text.) The tracts are ranked from lowest to highest on income as shown in Table 2b. If the tracts were to be divided into 2 groups, all tracts with ranks 1-6 would be in one group and all those with rank 7-12 would be in the other. Similarly if 3 groups were to be formed, the first group would contain tracts with ranks 1-4; the second group, ranks 5-8; and the third group, ranks 9-12.

Let us now plot the income of the 12 tracts and denote the partition into 2 and 3 groups by the parentheses around the x's representing income (Figures 2a and 2b). Two difficulties become clear here:

a. There is no indication what the optimum number of groups is.
b. Even if we assume that either 2 or 3 is the correct nunber of groups, the groups themselves do not appear to be "natural."

As an example of b., notice that in the partition into 2 groups, the tract with income of $11,000 appears to be distant from others members of its group. This difficulty is caused by the distortion of distances in the ranking process. A grouping that might appeal to our intuition is given in Figure 2c. (Notice that we intuitively pick the "correct" number of groups while at the same we determine group composition.) This grouping seems reasonable because distances between groups appear large with respect to distances between points in the same group. These distinctions disappear in ranking. The differences in income between tract 7 and 10 is $500 while that between tract 10 and 1 is $5,000. The difference in ranks, however, is 1 in each case. (Table 2b).

The method of summed-ranks is a simple extention of the method of ranking on one variable. Let p variables be measured on each census tract. The tracts are ranked on each variable separately, the p ranks are summed for each tract, and this sum is finally ranked.

EXAMPLE: Let each of 8 tracts have a median family income and median education of household head associated with it (Table 2c). Each tract can be plotted as a point in 2-dimensional space as shown in Figure 2d. Tracts 3 and 7 exhibit quite different behavior and are therefore distant from each other on the graph. A glance at Table 2c, however, reveals that they have the same rank. The difficulty here is that a 2-dimensional problem is being forced into 1 dimension. Although it was reasonable to order the tracts on each variable separately, there was no justification for ordering the tracts on both variables simultaneously. Only in the case where two variables are highly correlated is it valid to represent their ordering by the summed-rank.

The difficulties shown in this example occur in higher dimensions and

are compounded with the problem of distortion of distances, illustrated above in 1-dimension.

A reverse type of problem can also occur. Assume three variables are measured on each tract and that two are highly correlated. These two variables may be different names for the same phenomenon and yet they are treated as being independent. They are therefore given more weight than they are due in the method of summed-ranks.

3. Principal Components

The method of principal components often allows us to replace an initial set of variables with one index number. The method is demonstrated graphically for the 2-dimensional case. Let two variables x_1, x_2 be measured on each tract and plotted as in Figure 3a. An axis is drawn through the origin such that the sum of squares of the perpendicular distances of the points to the axis is minimized. This axis is called the principal component. The tract is now represented by one number: the distance from the origin of its projection on the principal component axis. This number takes the form $y = c_1x_1 + c_2x_2$, where the c's are known constants.

Representing each tract by its principal component value is justified only if the dispersion of points is primarily along the direction of the principal component axis. If this is the case, the tracts can be ranked and grouped on the basis of this value. This procedure gives rise to many of the same problems encountered in summed-ranks:

a. One index number can usefully replace the original variables only if the variation is primarily in the direction of the principal component axis. For this to happen the variables must be highly correlated.

b. If the principal component values are ranked to form groups, the problem of distortion of distance and the number of groups to consider again arises.

4. Cluster Analysis

The particular clustering technique applied here explores the structure of multivariate data in search of "clusters" by means of a certain criterion function. Each object has p variables associated with it and therefore can be represented by a point in p-dimensional space. The criterion function measures the ratio of the total dispersion of all points to the pooled dispersion of points within clusters. The goal is to find a grouping or clustering of points which maximizes the criterion function.

One-dimensional case

Consider the configuration of points $x_1, x_2, \cdots, x_6$ with the two possi- groupings shown in Figures 4a and 4b. The groups (or clusters) in 4a appear more compact, i.e. the dispersion or scatter of points within each group appears small with respect to the total scatter of all points.

Total scatter T is expressed mathematically as follows:

$$T = \sum_{i=1}^{6} (x_i - \bar{x}_T)^2 \text{ where } \bar{x}_T = 1/6 \sum_{i=1}^{6} x_i$$

The pooled-within groups scatter W is given by

$$W = W_1 + W_2$$

For Figure 4a,

$$W_1 = \sum_{i=1}^{4} (x_i - \bar{x}_1)^2 \text{ where } \bar{x}_1 = 1/4 \sum_{i=1}^{4} x_i$$

$$W_2 = \sum_{i=5}^{6} (x_i - \bar{x}_2)^2 \text{ where } \bar{x}_2 = 1/2 \sum_{i=5}^{6} x_i$$

For Figure 4b,

$$W_1 = \sum_{i=1}^{3} (x_i - \bar{x}_1)^2 \text{ where } \bar{x}_1 = 1/3 \sum_{i=1}^{3} x_i$$

$$W_2 = \sum_{i=4}^{6} (x_i - \bar{x}_2)^2 \text{ where } \bar{x}_2 = 1/3 \sum_{i=4}^{6} x_i$$

The criterion function is defined as the ratio T/W. Notice that T is constant under both groupings. Therefore maximizing T/W is equivalent to minimizing W. If the grouping in 4a is actually better than that in 4b then its value for W should be smaller. To find the optimum grouping into two clusters, all possible assignments of the points into two groups should be attempted until T/W is maximized.

In general the criterion function for the 1-dimensional case is defined as follows. Let x be a variable measured over each of n objects (here tracts).

Suppose the tracts are partitioned into g groups with the first group containing n_1 tracts with respective values $x_{11}, x_{12}, \cdots, x_{1n_1}$;

the second group containing n_2 tracts with respective values

$$x_{21}, x_{22}, \cdots, x_{2n_2};$$

the g-th group containing n_g tracts with respective values

$$x_{g1}, x_{g2}, \cdots, x_{gn_g}.$$

Then total scatter is given by

$$(1) \quad T = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij}-\bar{x}_T)^2 = n\sigma_T^2, \quad \sigma_T^2$$

is the variance of the entire collection of points;

$$\bar{x}_T = \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} x_{ij}$$

Pooled-within group scatter,

$W = W_1 + W_2 + \cdots + W_g$ where

$$(2) \quad W_i = \sum_{j=1}^{n_i} (x_{ij}-\bar{x}_i)^2 = n_i \sigma_i^2, \quad \sigma_i^2$$

is the variance of points in the i-th group;

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

Criterion function = T/W. All possible assignments of n points into g groups are attempted. The grouping which maximizes T/W is considered optimum.

Two-dimensional case

Let two variables x, y be measured over each tract and the tracts be partitioned into g groups as in the preceding paragraph. (The notation for the subscripts of the y's will be the same as that for the x's.) Then total scatter is given by the 2x2 determinant $|T|$

$$\text{where } T = \begin{bmatrix} n(\sigma_x)_T^2 & n[\text{cov}(x,y)]_T \\ n[\text{cov}(x,y)]_T & n(\sigma_y)_T^2 \end{bmatrix}$$

$$n[\text{cov}(x,y)]_T = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij}-\bar{x}_T)(y_{ij}-\bar{y}_T)$$

and $n(\sigma_x)_T^2$ is given by equation (1). The expression for $n(\sigma_y)_T^2$ is completely analogous.

Pooled-within group scatter is given by $|W|$ where

$$W = W_1 + W_2 + \cdots + W_g$$

$$\text{where } W_i = \begin{bmatrix} n_i(\sigma_x)_i^2 & n_i[\text{cov}(x,y)]_i \\ n_i[\text{cov}(x,y)]_i & n_i(\sigma_y)_i^2 \end{bmatrix}$$

$n_i(\sigma_x)_i^2$ is given by equation (2),

$n_i(\sigma_y)_i^2$ is calculated in the same way, and

$$n_i[\text{cov}(x,y)]_i = \sum_{j=1}^{n_i} (x_{ij}-\bar{x}_i)(y_{ij}-\bar{y}_i)$$

The criterion function is $|T|/|W|$.

It should be noted that the total scatter determinant

$$|T| = n^2\left\{(\sigma_x)_T^2 (\sigma_y)_T^2 - [\text{cov}(x,y)]_T^2\right\}$$

may be thought of as n^2[(length of scatter) x (width of scatter) - (overlap due to correlation)], i.e. the total area of scatter. Similarly $|W|$ can be considered the pooled-within group area of scatter. As in the 1-dimensional case, all possible assignments of the n points into g groups are attempted. The grouping which maximizes $|T|/|W|$ is considered the optimum.

The concepts presented above can be extended to any dimension p. In multivariate statistical theory, pxp determinants such as $|T|$ and $|W|$ (excluding the factor of n^2) are known as generalized variances and are often interpreted as representing volumes of dispersion.

EXAMPLE: Calculation of criterion function for 2-dimensional case. Consider the partitioning into two groups of the following points: (0,6), (2,12), (10,2), (12,4), (12-2), (14,4).

The criterion function will be calculated for two possible clusterings into two groups (see Figure 4c).

Clustering A: First group contains (0,6), (2,12).

$$\bar{x}_1 = 1, \quad \bar{y}_1 = 9$$

$$2(\sigma_x)_1^2 = (0-1)^2 + (2-1)^2 = 2;$$

$$2(\sigma_y)_1^2 = (6-9)^2 + (12-9)^2 = 18$$

$$2[\text{cov}(x,y)]_1 =$$

$$(0-1)(6-9) + (2-1)(12-9) = 3+3 = 6$$

$$W_1 = \begin{bmatrix} 2 & 6 \\ 6 & 18 \end{bmatrix}$$

Second group contains (10,2), (12,4), (12-2), (14,4).

$$\bar{x}_2 = 12, \ \bar{y}_2 = 2$$

$$4(\sigma_x)_2^2 = 8; \ 4(\sigma_y)_2^2 = 24;$$

$$4[\mathrm{cov}(x,y)]_2 = 4$$

$$W_2 = \begin{bmatrix} 8 & 4 \\ 4 & 24 \end{bmatrix}$$

$$W = W_1 + W_2 = \begin{bmatrix} 10 & 10 \\ 10 & 42 \end{bmatrix};$$

$$W = (10)(42) - 10^2 = 420 - 100 = 320$$

Clustering B: First group contains (0,6), (2,12) (10,2), (12,4)

$$\bar{x}_1 = 6, \ \bar{y}_1 = 6$$

$$4(\sigma_x)_1^2 = 104; \ 4(\sigma_y)_1^2 = 56;$$

$$4[\mathrm{cov}(x,y)]_1 = -52$$

$$W_1 = \begin{bmatrix} 104 & -52 \\ -52 & 56 \end{bmatrix}$$

Second group contains (12,-2), (14,4).

$$\bar{x}_2 = 13, \ \bar{y}_2 = 1$$

$$2(\sigma_x)_2^2 = 2;$$

$$2(\sigma_y)_2^2 = 18, \ 2[\mathrm{cov}(x,y)]_2 = 6$$

$$W_2 = \begin{bmatrix} 2 & 6 \\ 6 & 18 \end{bmatrix}$$

$$W = W_1 + W_2 = \begin{bmatrix} 106 & -46 \\ -46 & 74 \end{bmatrix};$$

$$|W| = (106)(74) - (46)^2 = 5728$$

For both clusterings:

$$\bar{x}_T = 8.33 \quad \bar{y}_T = 4.33$$

$$6(\sigma_x)_T^2 = 171.33; \ 6(\sigma_y)_T^2 = 107.33;$$

$$6[\mathrm{cov}(x,y)]_T = -92.67$$

$$|T| = \begin{vmatrix} 171.33 & -92.67 \\ -92.67 & 107.33 \end{vmatrix} = 9801.12$$

Clustering A: $|T| / |W| = 30.63$
Clustering B: $|T| / |W| = 1.71$

It was obvious from Figure 4a that A is a much better clustering than B. This has now been verified by the larger value of A's criterion function.

It is clear from the definition of the criterion function that the difficulties observed in the methods of summed-ranks and principal components have been eliminated. Distance is preserved by use of variances to measure dispersion, correlation between variables is accounted for by the covariance term in the scatter matrices, and the use of generalized areas or volumes rids us of the notion of strict ordering of objects.

The question of how many groups to take remains. Regardless of the number of groups taken, the total scatter remains the same. The pooled-within group scatter for the optimum grouping decreases, however, as the number of groups is increased. This, of course, causes an increase in $|T| / |W|$. Experience indicates that log $|T|/|W|$ tends to reach a plateau at a certain point, and an increase in the number of groups gives diminishing returns. The point at which the plateau begins is taken as the optimum number of groups.

5. A Clustering Computer Program

An IBM computer program employing the methods of section 4 has been written in G or H level FORTRAN and in 360 assembler language. (See Rubin and Friedman [2]. Some of the material in this reference is identical to that in [1]. The remaining material concerns other methods of clustering and instructions for utilizing the programs.) In addition to performing the computations directly related to the determination of clusters, the program produces auxiliary output which is necessary for a complete understanding of the clustering process. Two examples of this are the plot of tracts in eigenvector space and the calculation of discriminant weights.

In any classification problem it is not unusual to have objects which do not clearly belong to any group. In analysis of the census tracts of Washington, D. C., there were often tracts with values of certain variables which placed them far from the mean of any group. A plot of the tracts in a certain eigenvector space (see [1]) enables us to identify such outliers.

The discriminant weights indicate which variables play the greatest

role in distinguishing one cluster from another. A complete discussion of discriminant analysis can be found in Anderson [3] and Morrison [4].

Examples of eigenvector plots and discriminant weights are given in the next section.

6. Clustering Census Tracts of Washington, D. C.

Cluster analysis was applied separately to three different sets of variables measured on each census tract of Washington, D. C. The tracts composing each cluster were listed in the output of the program mentioned in Section 5. The mean value of every variable over each cluster was also computed. It is this set of mean values which characterizes the cluster.

Data for two of the sets of variables, "conditions surrounding birth" (1969) and welfare (1967), came from agencies in the District of Columbia government. The tracts in this case were based on 1960 census tract boundaries and were 122 in number.

The third set of variables was meant to serve as a general socio-economic indicator. The data and tract boundaries were taken from the 1970 Census. The 1970 tracts were 147 in number.

Table 6a gives the mean values of the clusters formed on the basis of five "conditions surrounding birth" variables for the year 1969. There is a clear ordering from BEST to WORST groups simultaneously on all variables. It is worth noting the unequal number of tracts in each group. This would not have occurred in the method of summed-ranks.

In practice the rule for determining the optimum number of groups is often quite vague. Exactly where the point of diminishing returns occurs in the criterion function is not always obvious. For this case, however, the choice seemed clear. Table 6b indicates that it is reasonable to take three groups.

The discriminant weights in Table 6c show that the percent of mothers under 20 years of age had a primary role although prematal care was also important in distinguishing between BEST and MEDIUM groups. Age of mothers was again dominant, although to a less extent, and both prenatal care and illegitimacy had secondary roles in distinguishing MEDIUM from WORST groups.

The next set of variables considered was the caseload, expressed as a percent of the population at risk, in each of four welfare categories (Table 6d). Here there is a high degree of skewness with the great majority of tracts belonging to the BEST group. Although the group in the second column is labelled MEDIUM, it is the worst in AFDC. This is a clear case where the tracts cannot be ordered on all variables simultaneously. Table 6e demonstrates the difficulty in choosing the optimum number of groups. Either 3 or 4 seemed appropriate.

Figure 6a is a plot of the tracts in eigenvector space as explained in [1]. The tracts in the BEST group are plotted as B's, etc. The BEST group appears more compact than the other groups. By means of other output from the computer program, it is possible to identify the outlying tract represented by the encircled "M" in the MEDIUM group and to determine which variables caused it to be so distant from its group mean.

The clusters formed in an analysis of four variables chosen as a general socio-economic indicator illustrate an interesting phenomenon (Table 6f). Although there is generally clear ordering of all variables from BEST to WORST, the distinction sometimes disappears, as in comparing the matriarchy and overcrowding indices between the POOR and WORST groups. The incomplete plumbing index seems to be the dominant variable in distinguishing between these groups. The discriminant weights in Table 6g verify this.

7. Additional Remarks

The method of clustering discussed in this paper is one of several which may be appropriate for classifying census tracts of a city. [2], for example, presents other methods and also considers various options to be used with the method applied here. In future studies, we plan to use a hybrid clustering model which will employ judgments from subject matter experts as well as mathematical techniques. One significant result of this will be the subjective weighting of variables before they are entered into the clustering process. At present all variables are assumed to be of equal importance.

REFERENCES

[1] Rubin, J. and Friedman, H. P., "On Some Invariant Criteria for Grouping Data," Journal of the American Statistical Association, December 1967.

[2] Rubin, J. and Friedman, H. P., "A Cluster Analysis and Taxonomy System for Grouping and Classifying Data," IBM Program Library, August 1967.

[3] Anderson, T. W., An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, Inc., 1958.

[4] Morrison, D. F., Multivariate Statistical Methods, McGraw Hill, Inc., 1967.

TABLE 2-a
Distribution of Income by Census Tract
(in $1,000 units)

Census Tract No.	1	2	3	4	5	6	7	8	9	10	11	12
Median Family Income	11.0	4.5	12.0	13.0	4.0	14.0	5.5	5.0	18.0	6.0	20.0	19.0

TABLE 2-b
Census Tracts Ranked by Income

Rank	1	2	3	4	5	6	7	8	9	10	11	12
Income	4.0	4.5	5.0	5.5	6.0	11.0	12.0	13.0	14.0	18.0	19.0	20.0
Tract No.	5	2	8	7	10	1	3	4	6	9	12	11

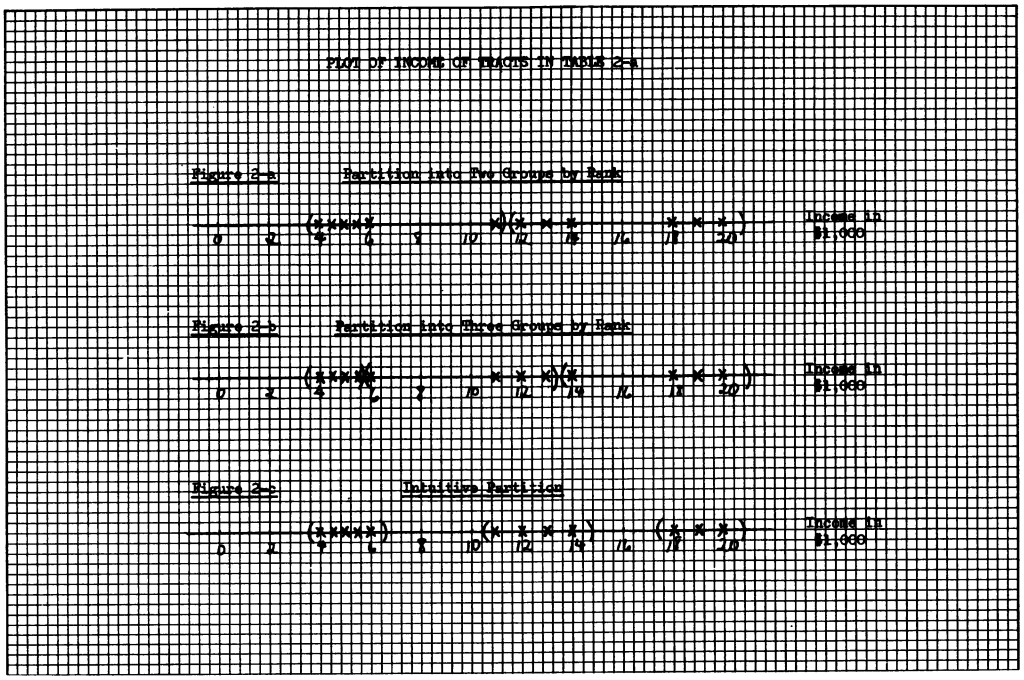
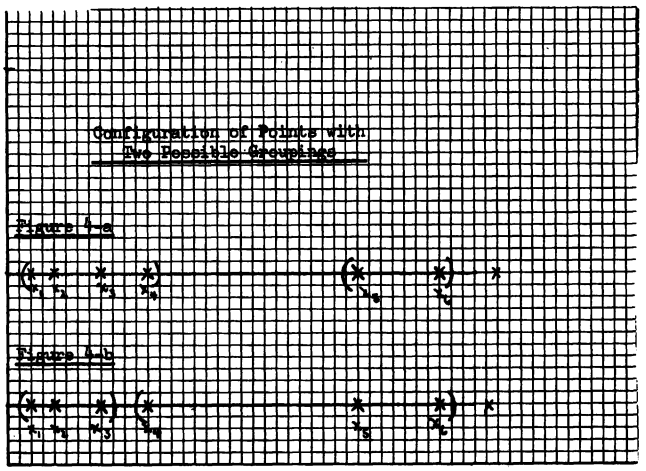
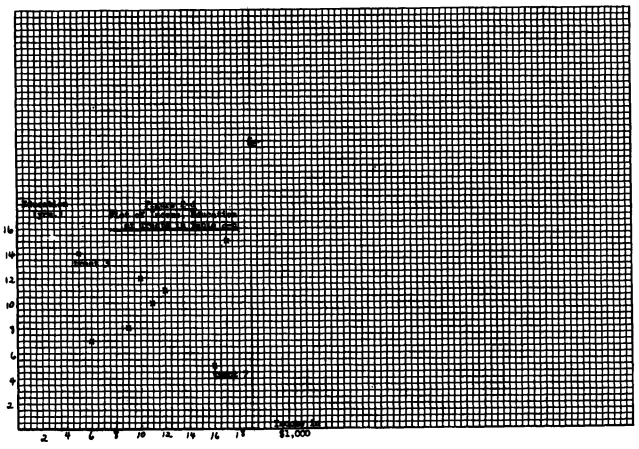


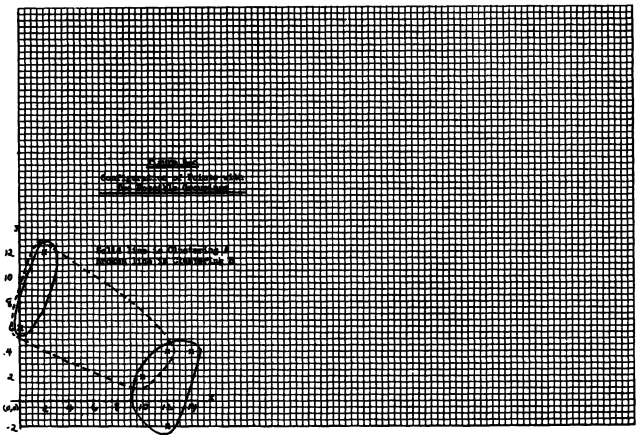
Table 2-c

APPLICATION OF SUMMED RANKS TO DISTRIBUTION OF INCOME AND EDUCATION OVER EIGHT TRACTS

Census Tract No.	1	2	3	4	5	6	7	8
Income ($1,000 units)	11	12	5	10	9	6	16	17
Education (yrs.)	10	11	14	12	8	7	5	15
Rank on Income	5	6	1	4	3	2	7	8
Rank on Education	4	5	7	6	3	2	1	8
Summed-rank	9	11	8	10	6	4	8	16
Final Rank	5	7	3½	6	2	1	3½	8







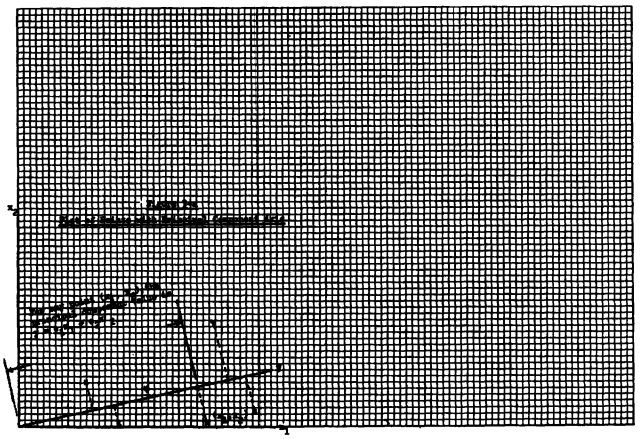


Table 6-a

CONDITIONS SURROUNDING BIRTH - 1969

	GROUP MEANS IN PERCENT*		
	BEST	MEDIUM	WORST
Mothers Under Age 20	5.2	23.2	38.8
No or Inadequate Prenatal Care	8.2	25.0	35.4
Birth Weight Under 5½ Lbs.	7.0	13.4	15.2
Illegitimate Births	12.9	27.4	51.0
Infant Mortality	2.1	3.2	3.2
Number of Tracts	30	51	41

* Base population is the total number of births.

Table 6-b

Maximum Value of Criterion Function by Number of Groups

NUMBER OF GROUPS	CRITERION FUNCTION	INCREMENT
2	1.25	
		1.21
3	2.46	
		.73
4	3.19	
		.45
5	3.64	
		.43
6	4.07	

Table 6-c

Discriminant Weights Between Groups

VARIABLES	DISCRIMINANT WEIGHTS	
	Best-Medium	Medium-Worst
Mothers Under Age 20	.78	.51
No or Inadequate Prenatal Care	.45	.29
Birth Weight Under 5-1/2 Lbs.	-.19	-.08
Illegitimate Births	-.002	.27
Infant Mortality	-.04	-.08

Table 6-d

WELFARE - 1967

PUBLIC ASSISTANCE CATEGORIES

GROUP MEANS IN PERCENT			
	BEST	MEDIUM	WORST
Old Age Assistance	2.7	7.7	9.0
Aid to Families with Dependent Children	3.4	23.2	12.0
Aid to Permanently and Totally Disabled	.6	1.5	3.1
General Public Assistance	.2	.3	.6
Number of Tracts	94	10	18

Table 6-e

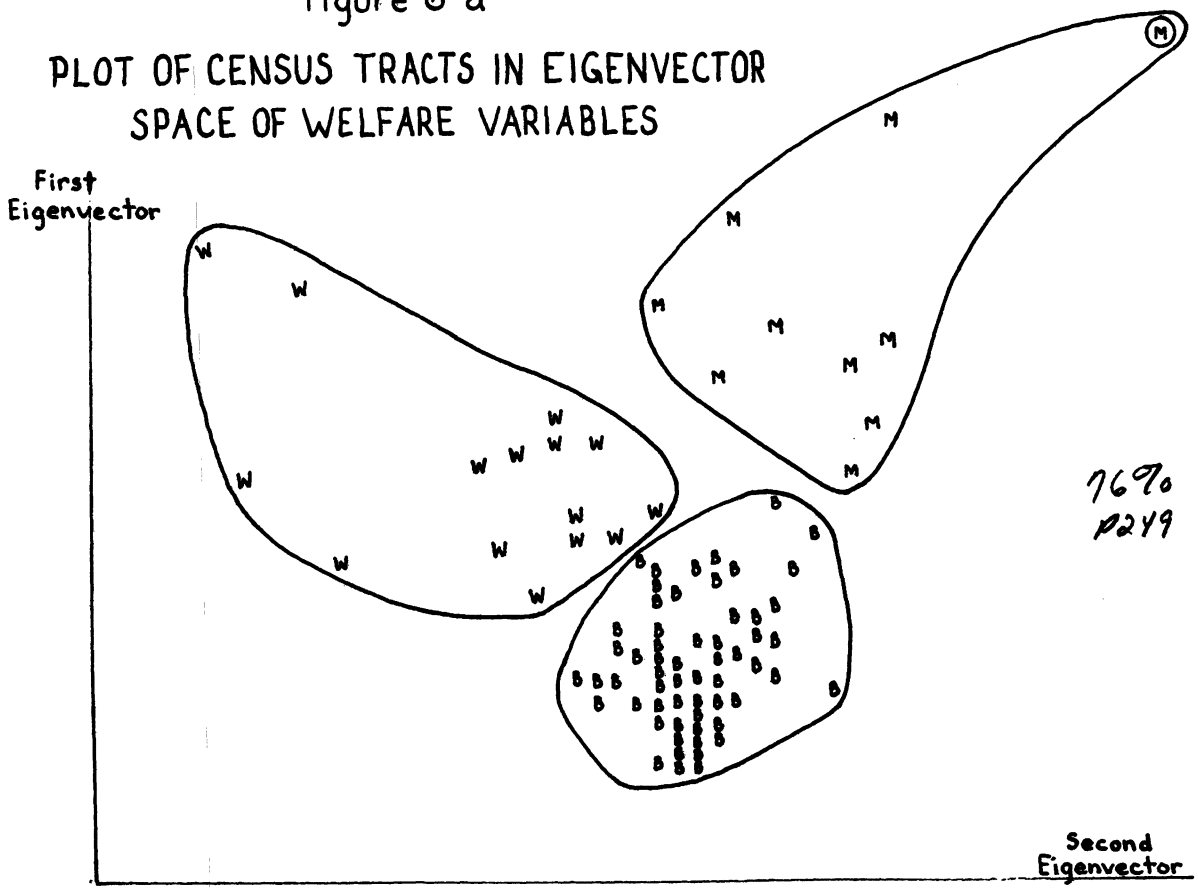
WELFARE - 1967

MAXIMUM VALUE OF CRITERION FUNCTION BY NUMBER OF GROUPS

Number of Groups	Criterion Function	Increment
2	1.22	
		.86
3	2.08	
		.86
4	2.94	
		.76
5	3.70	
		.76
6	4.46	

Figure 6-a

PLOT OF CENSUS TRACTS IN EIGENVECTOR SPACE OF WELFARE VARIABLES



769
P249

Table 6-f

SELECTED SOCIO-ECONOMIC VARIABLES

GROUP MEANS - 1970				
	BEST	MEDIUM	POOR	WORST
Median Family Income*	$17,000	$8,600	$6,700	$4,800
Matriarchy Index	15.5%	26.4%	38.4%	39.1%
Overcrowding Index	1.9%	8.8%	22.4%	21.4%
Incomplete Plumbing	.8%	2.2%	2.1%	16.8%
Number of Tracts	22	62	58	5

* This is not yet available from the 1970 Census. The values were estimated for each tract from a regression model developed by Westat Research, Inc., Rockville, Maryland.

Table 6-g

SELECTED SOCIO-ECONOMIC VARIABLES

VARIABLES	DISCRIMINANT WEIGHTS POOR - WORST
Median Family Income	- .03
Matriarchy Index	- .21
Overcrowding Index	.05
Incomplete Plumbing Index	1.10