Report Number 94-17

research

Transit System Monitoring and Design

Report Documentation Page

1. Report No.	2.	3. Recipient's Accession No.
MN/RC - 94/17		

4. Title and Subtitle	5. Report Date
Transit System Monitoring and Design	1990
	6.

7. Author(s)	8. Performing Organization Report No.
Yorgos J. Stephanedes, P.E.	

9. Performing Organization Name and Address	10. Project/Task/Work Unit No.
Civil & Mineral Engineering Department University of Minnesota 500 Pillsbury Dr. SE Minneapolis, Mn 55455	11. Contract(C) or Grant(G) No. (C) Mn/DOT 64953 TOC #37

12. Sponsoring Organization Name and Address	13. Type of Report and Period Covered
Minnesota Department of Transportation Office of Research Administration 200 Ford Building-Mail Stop 330 117 University Avenue St. Paul, Mn. 55155	Final Report 1988-1991
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract (Limit: 200 words)

Statistical techniques were developed for extracting the most significant features (indicators) from a transit system data base, and classifying proposed and existing transit systems according to the selected features. The data base was constructed by using information from all previous years available by the Mn/DOT, the Census and other sources to be used in classifying transit systems. The data base emphasized the use of raw characteristics of the operating system and the area socioeconomics. The feature extraction was done so that the minimum number of features were extracted that can be used for classifying the transit systems with maximum accuracy. The classification method was designed around the data base and is flexible so that it can use future data to update the data base at minimum cost. The transit system patterns, resulting from the classification method, were identified according to need and performance, and the main characteristics were specified for each pattern. These characteristics and descriptions identifying each pattern determines whether it should be modified. A controlled experiment was required to test the classification method. A randomly selected part of the data was classified by the method, and then the unselected data was treated as a control group for the experiment. After the experiment a percent of misclassifications was calculated.

17. Document Analysis/Descriptors	18. Availability Statement
Large-Stone Asphalt Mixtures Asphalt Mix Design Asphalt Mixture Gradation	No restrictions. This document is available through the National Technical Information Services, Springfield, Va. 22161

19. Security Class (this report)	20. Security Class (this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	159	

TRANSIT SYSTEM MONITORING AND DESIGN

FINAL REPORT

Principal Investigator: Yorgos J. Stephanedes, P.E.

Department of Civil and Mineral Engineering
University of Minnesota
Minneapolis, Minnesota

1990

CONTENTS.

LIST OF FIGURES.

Chapter 4 Page

Chapter 5 Page

LIST OF TABLES.

INTRODUCTION

1 Introduction.

Transit is an important component of the nation's total transportation system serving federal as well as state and local objectives. Public transit is often the only means of transportation to life-sustaining goods, services, and opportunities for many people, especially those who are elderly or mentally and physically handicapped.

The coordination of public transportation services is a concern of transit professionals at both the national and state levels. The continuing needs of public transit users, coupled with a decline in federal funds for transit, requires transit professionals to explore new ways of service delivery. Coordinating the use of existing transit services has the potential of improving service from both a cost and service efficiency perspective. Government agencies are increasingly mandating analysis of performance as a condition of financial aid.

Research of transit performance has been impeded by the absence of an acceptable classification that clusters similar systems together. Classification is a process which is basic to all sciences and which provides the earliest form of

measurement in a given discipline. More importantly, classification generates the concepts upon which a science can begin to build an understanding of the phenomena within its domain.

Government and private agencies are normally based on size to decide for funding a system but there have been no definitive studies that have specified and tested relationship between size and other operational variables, or with performance. Using peer groups for performance analysis addresses the controversial issue of whether transit systems should be compared. Transit managers tend to reject comparisons, yet most of them use comparative data for internal management assessments. Peer groups are typically selected based upon operating and service area characteristics.

Performance measures based on available operating, financial, and ridership statistics have recently been considered as criteria for the evaluation of public transit systems. Such measures can provide much insight into the operation of a particular system. In addition, these measures can be used to examine the differences among various transit systems and the changes that may occur from year to year. However, the injudicious application of generic performance indicators in the direct comparison of systems can provide misleading

information about the relative effectiveness of the system's operation and service. To compare systems adequately, it is necessary to adopt an approach that can allow for the unique local environments over which the operator has limited influence.

The generic term cluster analysis describes a large family of statistical classification procedures. Since the early 1960s, when high-speed computers made the use of this procedure relatively easy, more than 100 different clustering algorithms have been developed. Milligan (1981) conducted a computerized search of the literature in 1976 and showed that new or considerably revised algorithms were appearing at a rate of about one per month. Only few authors (e.g. D' Andrade 1978, von Eye and Wising 1978, 1980) however, have tried to compare their new procedures with already existing clustering algorithms.

Over the last years, there has been a growing interest in quasi-statistical techniques for forming classifications. These techniques are known under the generic name "cluster analysis". There are a large number of different cluster analysis methods, most of which fall into two families: hierarchical agglomerative methods and iterative partitioning methods. Hierarchical agglomerative methods start the

clustering process by forming a matrix which represents the pairwise similarities of all entities being clustered. These methods they proceed to build clusters gradually (i.e. to agglomerate) by putting the most similar entities together. This agglomerative process can be represented by hierarchical trees or dendrograms. The second major family of clustering methods are the iterative partitioning methods. The methods begin with a predetermined classification (i.e. partition) and through various iterative processes try to find a revised classification which will optimize a measure of homogeneity of the cluster.

In view of the diversity of available algorithms, the potential consumer of cluster analysis faces several problems. First because of that development of different terminologies in different fields of application, several labels are often used for the same clustering algorithm (Blashfield and Aldenderfer, 1978). More important few guidelines are available for choosing a clustering procedure for research applications. This problem is especially perplexing since different algorithms are likely to produce different solutions when applied to the same data set (Bartko, Straus and Carpenter, 1971). Furthermore, there is no guarantee that any of the available clustering algorithms will recover the true cluster structure either under error-free or error-perturbated

situations (Milligan 1980).

The objective of this work is to develop classification methods that give satisfactory results for transit research. A clustering method is introduced that does not have the disadvantage of the iterative partitioning methods, to predefine a classification. Usually this is a difficult task, and the user in many cases is unable to come up with a correct initial classification. This work is part of a larger project that seeks to provide improved methods for design of transit systems based on needs and the characteristics of the service area. Results of this research can be used as a descriptive framework for comparative studies. The orientation has been to provide a technique that would be useful for internal decision making within each transit agency.

LITERATURE REVIEW

2 Literature Review.

2.1 Previous Applications of Classification in Transportation.

Separating transit systems into peer groups that share similar operating characteristics is analogous to separating any set of objects into a number of groups in which members of the same group are more similar to each other than to objects in other groups, and the groups differ from one another. Problems of this sort are common in the social and biological sciences and in applied settings like market research. Tardif et al. (1977) clustered neighborhoods together to designate transit market areas for Sacramento, California. Other transportation researchers have used clustering techniques to facilitate research. Bottiny and Coley (1967) grouped urbanized areas for transportation analysis, and Golob et al. (1972) used similar procedures to group metropolitan areas for their analysis of arterial transportation requirements.

The principal use of the classification of transit systems is to assist comparison of performance between similar systems. In some studies group of data with different characteristics were used to define parameters of some equations. For example

Nelson (1972) and Veatch (1973) estimated supply and demand equations that are widely used. Nelson estimated his parameters using transit systems in urban ares, 51 systems in 1968 and 44 in 1960, although these systems were quite different in other operating characteristics. Veatch restricted his analysis to 29 systems operating in small and medium sized cities, in an attempt to control for differences in the operating environment. These studies should be replicated with systems drawn from one or more peer groups, after applying a classification technique, as it may significantly improve estimation results.

Another research area that could be improved by using the results of peer group analysis are current studies of the effect of subsidies on transit performance. Purcher et al. (1983) used a national sample of 77 systems in 1979 and 135 systems in 1980 for which reliable data were available. However this study included systems with quite different operating characteristics. Such studies would result in a wider acceptance, of their results if systems that are relatively homogeneous in operating characteristics, were used.

A national study have used size as the differentiating characteristic for comparative studies. The National Urban

Mass Transportation Statistics for fiscal years 1979 and 1980 grouped systems by the total number of revenue vehicles (U.S. Department of Transportation, 1981, p.vi). Research by Anderson and Fielding (1982) used three performance indicators to cluster transit systems. This method was replicated in a research, but was rejected when it was found that the clusters based upon size, peak-to-base operating ratio, total revenue vehicle miles and speed yielded superior results. The peer groups based on performance were neither as distinct from each other as those based on operating characteristics, nor did they capture as much of the variability of all seven performance indicators.

Later study from Fielding (1985), University of California proposed these four factors to reflect service area characteristics that constrain the decisions made by the transit managers. The hierarchical clustering technique was used to partition fixed route motor bus transit systems from different states of the United States. In this study Fielding was trying to classify the systems so that the separate clusters should capture the important difference between types of transit systems. Also the number of systems in each group should be sufficient for comparative analysis within the group but not so numerous that the task of comparison is excessive. The clusters should be formed at approximately the same level

in the cluster hierarchy so that they reflect about the same degree of similarity within each cluster. The clustering analysis resulted in twelve peer groups. Generally, there is more variability in the number of peak vehicles and vehicle miles and less variability in the speed and peak-to-base ratio. Some of the groups could have been divided into smaller cluster to reduce the variability, but this would have resulted in an undesirable large number of peak groups. The validity of these groups was examined by comparing the groups based not only on the four factors, but also on performance indicators. Statistically significant differences were found between the groups in terms of both the operating and the performance characteristics.

In another study at Purdue University (Sinha, 1985) a cluster analysis was done on Indiana transit systems. As with the previous study, factors beyond the control of the managers were chosen. The list of indicators came from a literature search and suggestions from the managers. A statistical correlation analysis was implemented for clustering the systems. First, a subset was admitted to a cluster if its correlation with a member of that cluster was greater than 0.81. Then, each subject was compared to the remaining members. Only subsets with an average correlation greater than 0.64 were retained in the group. The analysis performed

on data from 1981, 1982, and 1983 with consistent results. Four groups were found large, medium, small and all demand responsive systems. It was determined that most clusters differed in the factors at a significant level, the exception to this was the factor speed.

The last two clustering applications are more similar with this study. The University of California study examines systems throughout the United States and investigates only urban fixed-route systems. A large number of systems is included resulting in 12 pear groups. The research at Purdue University is probably more similar to this project. All transit systems in Indiana were analyzed, and 4 groups were found. The methods used in both analyses are valid, but this is not known whether the algorithms are the best for their objective. In this study the Karhunen-Loève extraction was used to reduce the dimensionality of the data. This was not done in the California and Indiana studies. Instead a smaller group of variables was initially selected but having as result the lost of information.

2.2 Literature on Classification Methods.

One correlate of the explosion of interest in cluster analysis

is the proliferation of various methods and software programs for performing cluster analysis. Each year there are a number of articles which advocate a new method of cluster analysis. The accumulation of these articles has been led to a large collection of diverse methods, all of which are subsumed under the generic title of cluster analysis.

There are two major types of cluster analysis methods: hierarchical agglomerative and iterative partitioning. These two classes of cluster analysis methods represent approaches which are hard to compare directly.

The simple cluster seeking algorithm (Tou, 1974) is a quick and simple algorithm. It can give useful results if the data is composed of distinctly separated clusters. This, unfortunately, is usually not the case. If the user has an idea of how many groups there are and where their centers are, the k-means algorithm (Tou, 1974) would be desirable. If there are only two variables, it is easy to view the points and pick out possible clusters. However, as the dimensions increase, this gets more difficult. The maximum distance (Tou, 1974) and decision directed (Young, 1974) algorithms do not require the user to predetermine the number of clusters, but instead require an idea of the distance between clusters. These algorithms are most useful when the desired distance

between clusters is known, and there is no restriction on the number to be found. Hierarchical clustering gives the user the option to decide how many clusters and how far apart they are after the analysis. Different cutoff levels can be tested for some optimum after analysis. The ISODATA algorithm (Tou, 1974) is more complex. However, the interactive nature of this algorithm makes it attractive. After each iteration the user can observe the data and make desired changes in a number of parameters if necessary. An algorithm given by Watanabe (1985) has the objective of minimizing the entropy of a data set. The algorithm is called dynamic coalescence model and can be thought of as a dynamic hierarchical clustering algorithm.

Each class of classification methods contains a large number of algorithms which have somewhat different properties. There are different hierarchical agglomerative methods. Sneath and Sokal (1973) discuss single linkage, complete linkage and four types of average linkage. In addition there are at least three different methods which attempt to minimize the variance within clusters (Anderberg, 1973; Ward, 1963) two methods concerned with optimizing information statistics for cluster structures (Clifford and Stephenson, 1975), two methods by McQuitty (1967) which are variations on the complete and average linkage methods as well as another variation on

complete linkage by Carlson (1972), and finally Lance and Williams (1967a, 1967b) have advocated a generic hierarchical agglomerative method which is based upon their theoretical overview of these methods (flexible theta method).

In addition to the hierarchical agglomerative methods, Hierarchical divisive methods have also been proposed in the biological sciences (Edwards and Cavalli-Sforza, 1965) and have been used in ecology (Clifford and Stephenson, 1975) and anthropology (Peebles, 1972; Whallon, 1972). Iterative partitioning methods are generally the focus in the pattern recognition literature and in statistics (Bezdek, 1974a, 1974b; Ball, 1965; Friedman and Rubin, 1967). Factor analysis has been a popular topic in multivariate statistics in psychology, hence psychologists have proposed a number of clustering methods whose topic is related to that of factor analysis (Tryon, 1939; Tryon and Bailey, 1970; Lorr and Badhakrishnan, 1967). Clumping (Peay, 1975), mode searching (Jones, 1968; Wishart, 1969) and graphic theoretic methods have also been proposed but have not stimulate a large literature.

Given the large number of cluster analysis methods and the diverse approaches to forming classifications which these methods represent, a natural question concerns which methods

are most frequently used. By looking the literature, nearly three of every four use one of the hierarchical agglomerative methods. There are three possible reasons why hierarchical agglomerative methods are being used so extensively:

- These methods are among the oldest of those available, and it is these methods which were popularized by the book of Sokal and Sneath (1963).
- Researchers tend to use whatever methods have been previously used in their literature; hence methods with a long history are likely to become dominant.
- Hierarchical agglomerative methods are the only methods which have been the object of empirical analysis.

Since the characteristics of the other methods are less well understood, researchers tend to stick to the hierarchical agglomerative methods.

Very little work has been done on the empirical evaluation of the properties of different clustering methods. The few studies which have been published have reached contradictory conclusions. Some studies have favored single linkage hierarchical agglomerative clustering (Fisher and van Ness, 1971; Jardine and Sibson, 1971; Sibson, 1971) while others

have favored complete linkage (Baker and Hubert, 1975; Bartko, Strauss and Carpenter, 1971). Still others have argued in favor of average linkage (Cunningham and Ogilive, 1972; Sneath, 1966; Sokal and Rohlf, 1962). And Ward's approach to minimum variance hierarchical agglomerative clustering also has its adherents (Blashfield, 1976a, 1976b)

Much of the confusion about the comparison of the properties of different methods of cluster analysis stems from the use of different theoretical and methological orientations for judging what forms an acceptable classification. For example, studies which favor average linkage all compare the dendrogram generated by his method to the structure of the similarity matrix which was used to start the clustering process. Those studies which favor single linkage, however, compared methods in terms of theoretical criteria which all clustering method should satisfy. Not surprisingly, the studies within each of the conceptual approaches agree on the clustering technique which they favor. Nevertheless, it is striking that even the studies which have attempted to consolidate our understanding of cluster analysis there is so much fragmentation and diversity.

2.3 Classification Methods.

It is evident that the ability to determine characteristics prototypes or cluster centers in a given set of data plays a central role in the design of pattern classifiers based on the minimum-distance concepts. The methods discussed constitute a cross section of representative approaches to the cluster-seeking problem. Cluster seeking algorithms are experiment-oriented techniques in the sense that the performance of a given algorithm is not only dependent on the type of data being analyzed, but is also strongly influenced by the chosen measure of pattern similarity and the method used for identifying clusters in the data.

2.3.1 Measures of Similarity.

To define a data cluster, it is necessary to first define a measure of similarity which will establish a rule for assigning patterns to the domain of a particular cluster center. The Euclidean distance between two patterns x and z is defined as:

$$D=\|x-z\| \qquad (2.3.1)$$

as a measure of their similarity - the smaller the distance, the greater the similarity. Another meaningful distance measure is the Mahalanobis distance from x to m:

$$D=(x-m)'C^{-1}(x-m) \qquad (2.3.2)$$

which is a useful measure of similarity when statistical properties are being explicitly considered. In the previous equation C is the covariance matrix of a pattern population, m is the mean vector, and x represents a variable pattern. The measures of similarity need not be restricted to distance measures. For example, the nonmetric similarity function:

$$s(x,z)=\frac{x'z}{\|x\|\|z\|} \qquad (2.3.3)$$

which is the cosine of the angle between the vectors x and z, is maximum when x and z are oriented in the same direction with respect to the origin. This measure of similarity is useful when cluster regions tend to develop along principal axes. On Fig. 2.3.1 the pattern z_1 is more similar to x than pattern z_2 since $s(x,z_1)$ is greater than $s(x,z_2)$.

However, the use of this similarity measure is governed by certain qualifications, such as sufficient separation of cluster regions with respect to each other as well as with respect to the coordinate system origin.

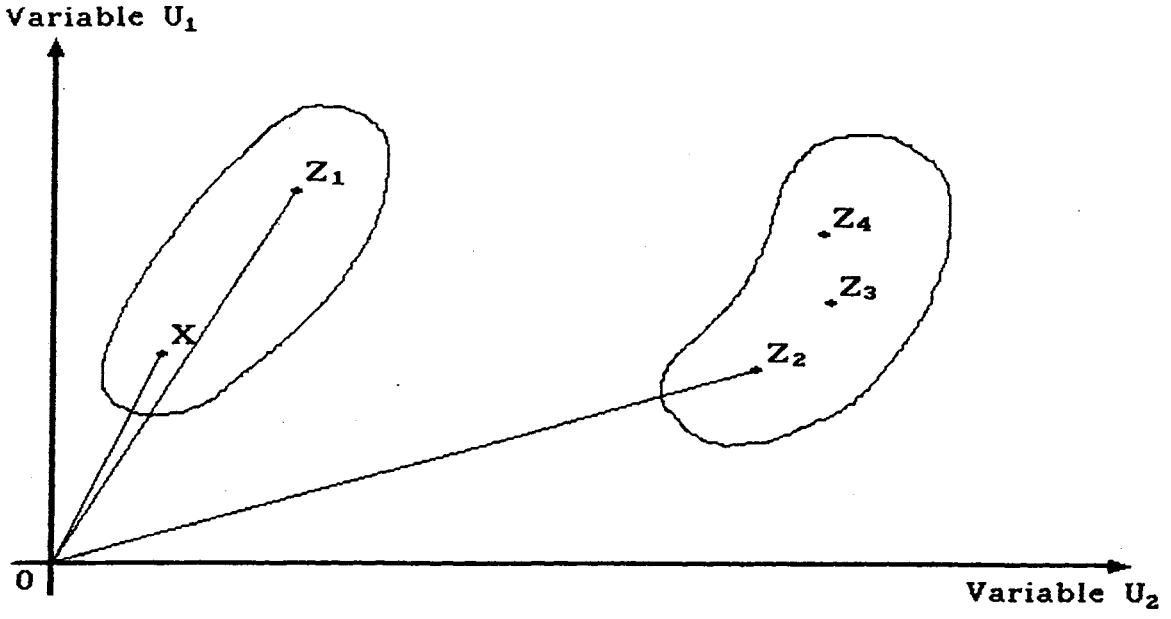

Figure 2.3.1 Illustration of a similarity measure.

2.3.2 Clustering Criteria.

After a measure of pattern similarity has been adopted, we have to specify a procedure for partitioning the given data into cluster domains. The clustering criterion used may represent a heuristic scheme, or it may be based on the minimization (or maximization) of a certain performance index.

The Euclidean distance measure readily lends itself to this approach because of its familiar interpretation as a measure of proximity. However, since the proximity of two patterns is a relative measure of similarity, it is usually necessary to establish a threshold in order to define degrees of acceptable similarity in the cluster-seeking process.

The performance-index approach is guided by the development of a procedure which will minimize or maximize the chosen performance index. One of the most often used indices is the sum of the squared errors index, given by:

$$J = \sum_{j=1}^{N_c} \sum_{x \in S_j} \|x - m_j\|^2 \qquad (2.3.4)$$

where N_c is the number of cluster domains, S_j is the set of samples belonging to the jth domain, and

$$m_j = \frac{1}{N_j} \sum_{x \in S_j} x \qquad (2.3.5)$$

is the sample mean vector of set S_j. N_j represents the number of samples in S_j. The index of Eq.(2.3.4) represents the overall sum of the squared errors between the samples of a cluster domain and their corresponding mean. There are numerous performance indices in addition to the previous one. Other common indices are the average squared distances between samples in a cluster domain, the average squared distances between samples in different cluster domains, indices based on the scatter matrix concept, and minimum- and maximum-variance indices.

2.3.3 A Simple Cluster-Seeking Algorithm.

Suppose that we have a set of N sample patterns $\{x_1, x_2, \ldots, x_N\}$. Let first cluster center z_1 be equal to any of the sample

patterns, and select an arbitrary nonnegative threshold T. Let choose $z_1 = x_1$. Next, we compute the distance D_{21} from x_2 to z_1. If this distance exceeds T, a new cluster center, $z_2 = x_2$, is started. Otherwise, we assign x_2 to the domain of cluster center z_1. Suppose that $D_{21} > T$ so that z_2 is established. In the next step, the distances D_{31} and D_{32} from x_3 to z_1 and z_2 are computed. If both D_{31} and D_{32} are greater than T, a new cluster center, $z_3 = x_3$, is created. Otherwise, we assign x_3 to the domain of the cluster center to which it is closest. In a similar fashion, the distance from each new pattern to every established cluster center is computed and thresholded, and a new cluster center is created if all of these distances exceed T. Otherwise, the pattern is assigned to the domain of the cluster center to which it is closest.

The results of the foregoing procedure depend on the first cluster center chosen, the order in which the patterns are considered, the value of T, and the geometrical properties of the data. The effects are illustrated in Fig. 2.3.2, where three different cluster center arrangements have been obtained for the same data simply by varying T and the starting point.

2.3.4 Maximin-Distance Algorithm.

The maximin (maximum-minimum) -distance algorithm is another simple heuristic procedure based on the Euclidean distance

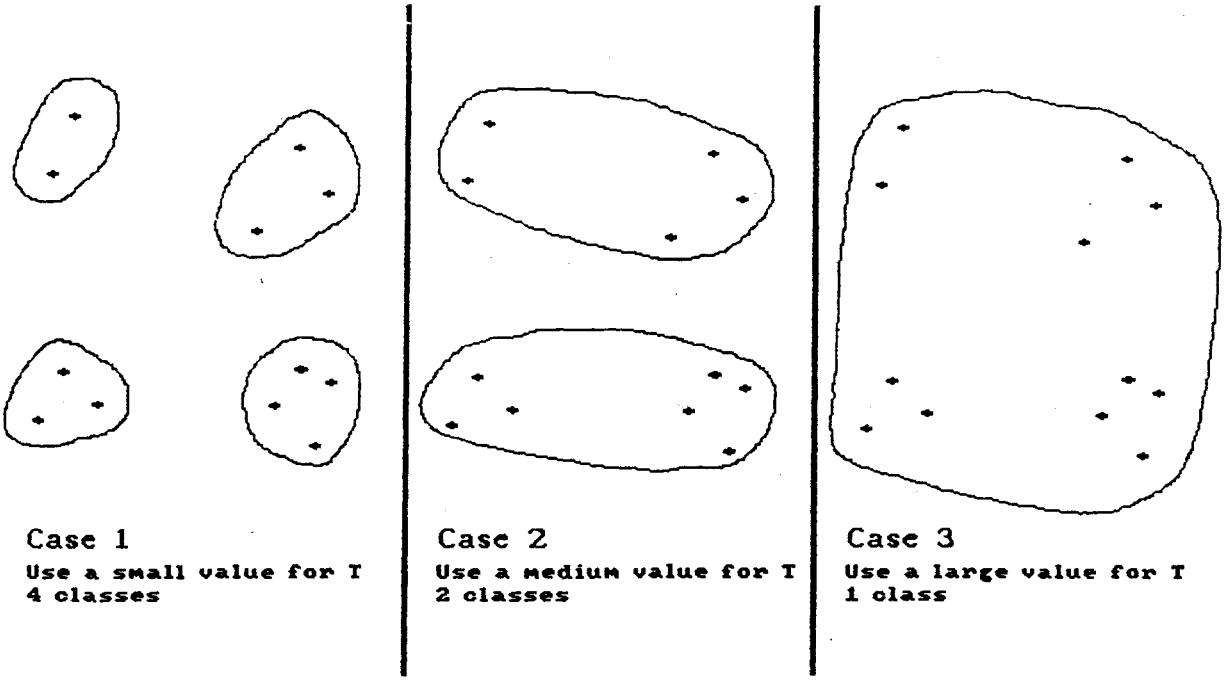

Figure 2.3.2 Effects of the threshold and starting points in the simple-seeking algorithm.

concept. This algorithm is similar in principal to the scheme of the simple cluster-seeking algorithm, with the exception that it first identifies the cluster regions which are farthest apart.

Suppose that we have a set of N sample patterns $\{x_1, x_2, \ldots, x_N\}$. This procedure requires an initial starting point. Additional points to be used as cluster centers are found by finding the furthest points from the first and subsequent cluster centers. After the first cluster center is given, the point that is farthest away (or least similar) from the first cluster center becomes the second cluster center. Subsequent

cluster centers are found by computing the distances from each of the remaining sample points to the cluster centers and saving the minimum of these distances for each sample point. The sample point whose minimum distance to the cluster centers is the largest is then considered. If this distance is greater than a threshold value, it becomes a new cluster center. Otherwise, this part of the algorithm ends, and all remaining sample points are assigned to the class of their nearest cluster center. The threshold value is often defined as a fraction of the average distance between cluster centers, 0.5 is typically used. A summary of the maximum distance algorithm is given below.

Step 1. Choose $z_1 = x_i$, where x_i can be any sample point.

Step 2. z_2 is the farthest sample from z_1.

Step 3. For z_3 to z_N:

Step 4. Compute the distance from each remaining sample, x_i, to all z's. Save the minimum of these distances for each x_i.

$$MIN_i = \min\{|x_i - z_j|, \ \forall j\} \qquad (2.3.6)$$

Step 5. Consider the sample with the largest minimum distance.

$$MAX\text{-}\max\{MIN_i, \forall i\} \qquad (2.3.7)$$

If that distance is an appreciable fraction of typical distances between existing z's, then the sample is a new cluster center. Go to step 4. Otherwise go to step 6.

Step 6. Assign the remaining samples to the cluster of the nearest z_j.

The maximum distance algorithm is illustrated in Fig. 2.3.3. In this simple example we have obtained three cluster centers x_1, x_6, and x_7. To assign the remaining samples to the domain of these centers we simply assign each sample to its nearest cluster center. Thus, we obtain the cluster domains $\{x_1, x_3, x_4\}$, $\{x_2, x_6\}$, and $\{x_5, x_7, x_8, x_9, x_{10}\}$. These results agree with the cluster domains that we would intuitively expect to get from these data.

The results of the maximin distance algorithm depend on the initial cluster center and the threshold value. The threshold value determines the number of classes. As the value of the threshold increases the number of resulting classes is decreased.

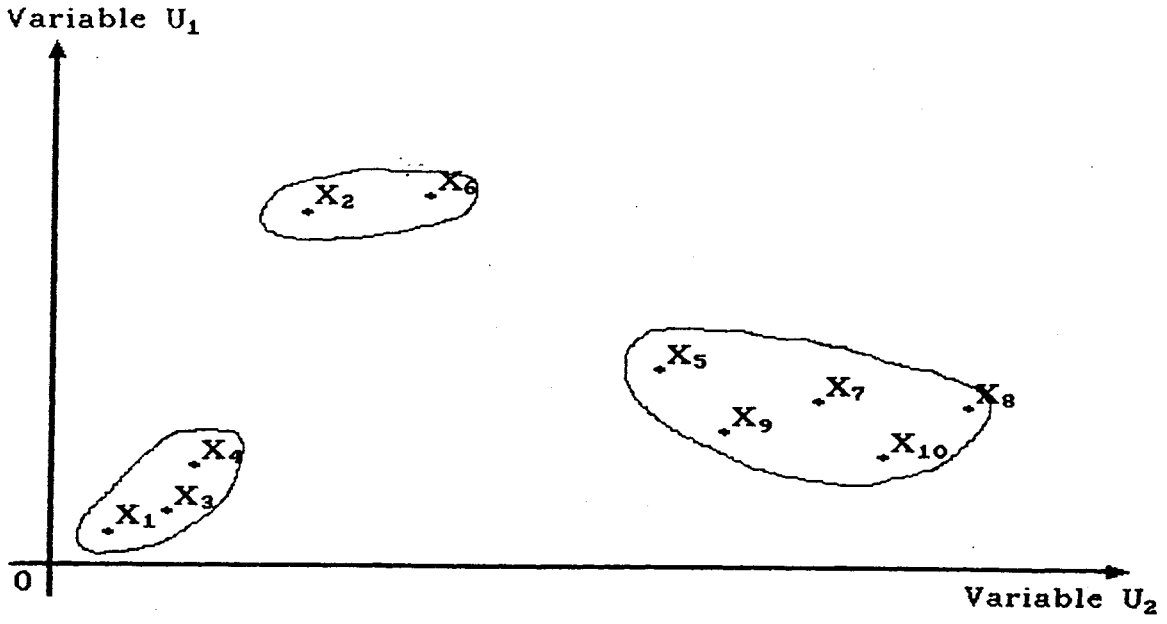

Figure 2.3.3 Sample patterns used in illustrating the maximum-distance algorithm.

2.3.5 K-Means Algorithm.

The K-Means algorithm is based on the minimization of a performance index which is defined as the sum of the squared distances from all points in a cluster domain to the cluster center. This procedure consists of the following steps.

Step 1. Choose K initial cluster centers $z_1(1)$, $z_2(1),\ldots,$ $z_K(1)$. These are arbitrary and are usually selected as the K samples of the given sample set.

Step 2. At the Kth iterative step distribute the samples {x} among the K cluster domains, using the

relation:

$$x \in S_j(k) \quad if \quad \|x - z_j(k)\| < \|x - z_i(k)\|, \quad \forall i = 1, 2, \ldots, K, \; i \neq j \qquad (2.3.8)$$

where $S_j(k)$ denotes the set of samples whose cluster center is $z_j(k)$. Ties in expression (2.3.8) are resolved arbitrarily.

Step 3. From the results of step 2, compute the new cluster centers $z_j(k+1)$, $j = 1, 2, \ldots, K$, such that the sum of the squared distances from all points in $S_j(k)$ to the new cluster center is minimized. In other words, the new cluster center $z_j(k+1)$ is computed so that the performance index:

$$J_i = \sum_{x \in S_j(k)} \|x - z_j(k+1)\|^2, \quad j = 1, 2, \ldots, K \qquad (2.3.9)$$

is minimized. The $z_j(k+1)$ which minimizes this performance index is simple the sample mean of $S_j(k)$. Therefore, the new cluster center is given by:

$$z_j(k+1) = \frac{1}{N_j} \sum_{x \in S_j(k)} x, \qquad (2.3.10)$$

$$j=1,2,\ldots,K$$

where N_j is the number of samples in $S_j(k)$. The name "K-Means" is derived from the manner in which cluster centers are sequentially updated.

Step 4. If $z_j(k+1) = z_j(k)$ for j = 1, 2,..., K, the algorithm has converged and the procedure is terminated. Otherwise go to step 2.

The behavior of the K-Means algorithm is influenced by the number of cluster centers specified, the choice of initial cluster centers, the order in which the samples are taken, and the geometrical properties of the data (Fig. 2.3.4).

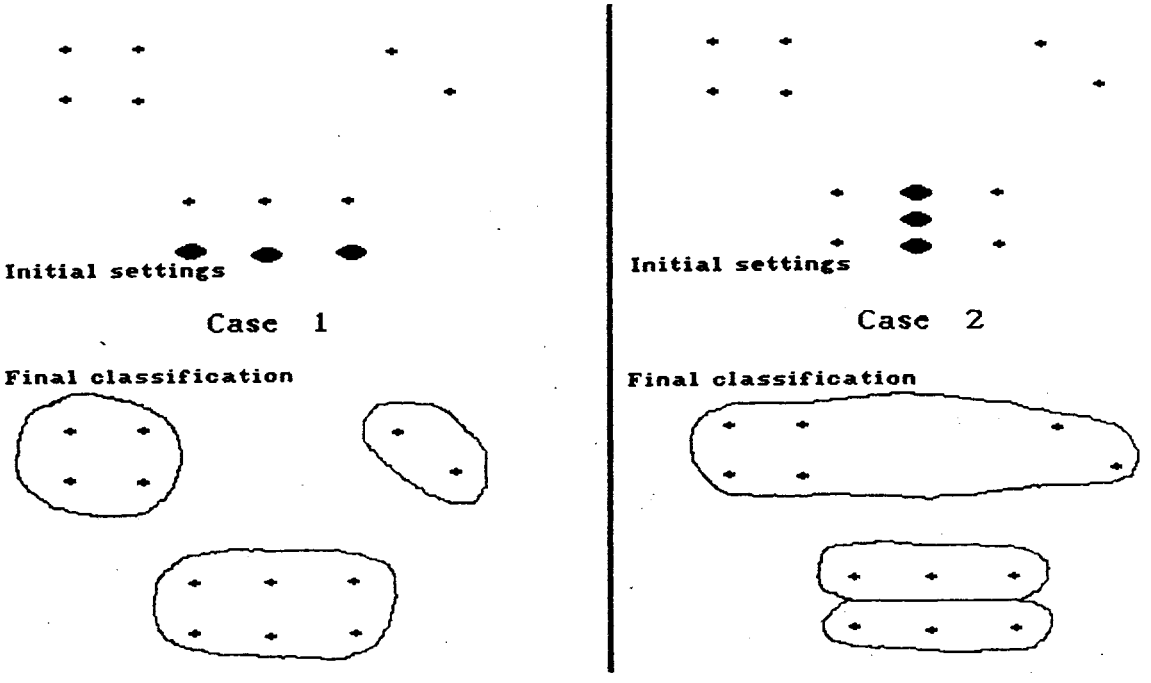

Figure 2.3.4 Illustration of k-means algorithm.

Although nogeneral proof of convergence exists for this algorithm, it can be expected to yield acceptable results when the data exhibit characteristic pockets which are relatively far from each other. In most practical cases the application of this algorithm will require experimenting with various values of K as well as different choices of starting configurations.

2.3.6 Hierarchical Clustering.

Hierarchical clustering begins with each sample point as a separate cluster. By using a measure of similarity, clusters are joined in a step by step process. Clusters that are most similar are first grouped together to form a new cluster. Then the similarity threshold is changed and clusters that are more similar than the threshold are grouped. This continues until all objects are in the same cluster Fig. 2.3.5). The results can be evaluated at a particular threshold value or at a desired value of k, the number of classes. The general algorithm follows.

Step 1. Begin with each single object, x_i, as a separate cluster z_i.

Step 2. Compute D_{ij} = the distance from z_i to z_j for all z's.

Step 3. If $D_{ij} < T$, merge z_i and z_j.

Step 4. Increase T and go to step 2 until all

clusters are merged into one.

In hierarchical clustering, when two entities are merged, the merger is permanent. This reduces the number of possibilities that need to be tested (compared to complete enumeration).

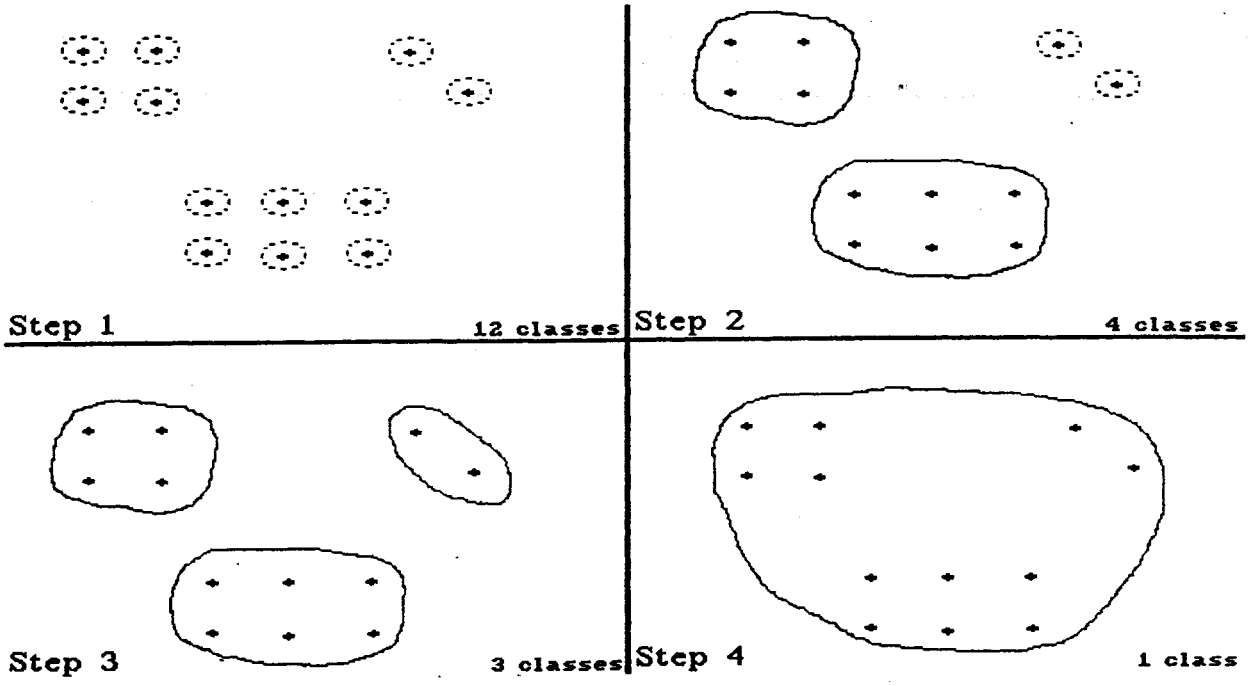

Figure 2.3.5 Illustration of hierarchical algorithm.

Several methods, or criteria, for forming clusters can be used. The most frequently used are:

- Single linkage.

  Distance is determined by the nearest neighbors of each cluster (Fig. 2.3.6a).

$$D_{tr} = \min(D_{pr}, D_{qr}) \qquad (2.3.11)$$

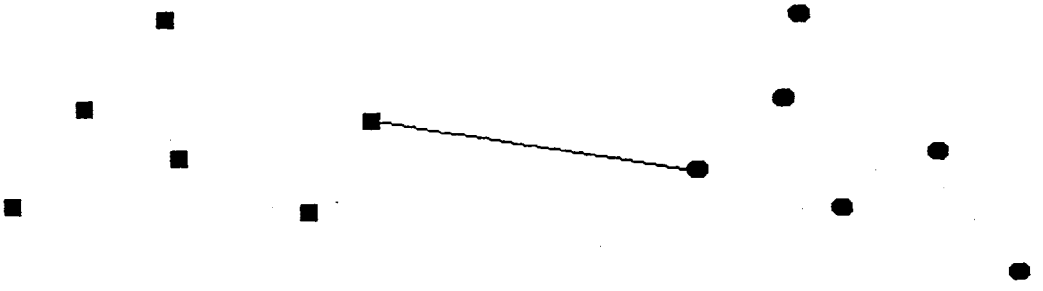

Figure 2.3.6a Simple linkage.

♦ Complete linkage.

Distance is determined by the furthest neighbors of each cluster (Fig. 2.3.6b).

$$D_{tr}=\max\{D_{pr}, D_{qr}\} \qquad (2.3.12)$$

♦ Centroid linkage.

The clusters with the most similar mean vectors are merged (Fig. 2.3.6c).

$$S_{tr}=\frac{N_p}{N_p+N_q}S_{pr}+\frac{N_q}{N_p+N_q}S_{qr}-\frac{N_pN_q}{N_p+N_q}S_{pq} \qquad (2.3.13)$$

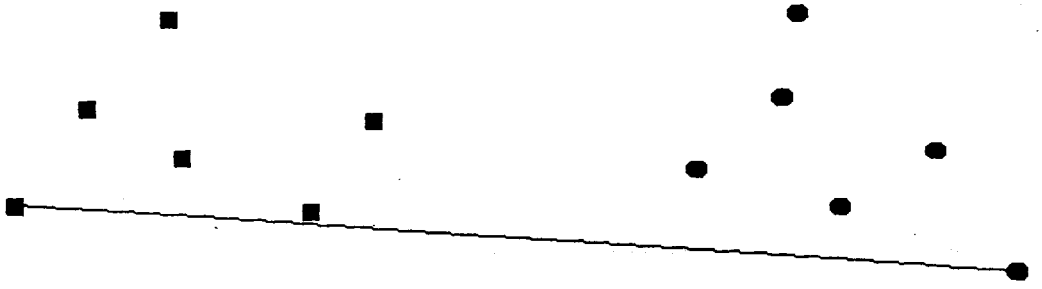

Figure 2.3.6b Complete linkage.

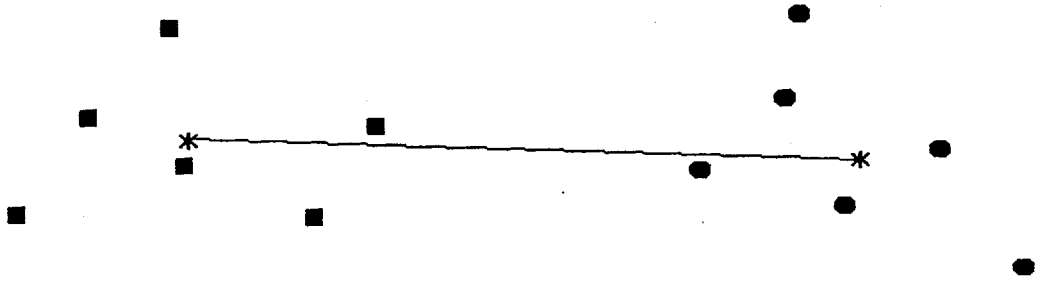

Figure 2.3.6c Centroid linkage.

The Squared Euclidean Distance is used.

♦ Average linkage between groups.

The average similarity between all entities in the new cluster is considered (Fig. 2.3.6d).

$$S_{tr}=S_{pr}+S_{qr} \qquad (2.3.14)$$

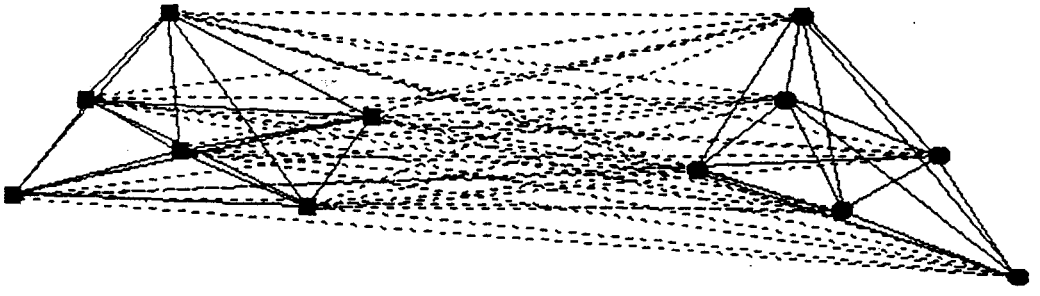

Figure 2.3.6d Average linkage between groups.

The average within group similarity for a cluster formed by merging t and r is:

$$\frac{SM_t+SM_r+S_{tr}}{(N_t+N_r)\,(N_t+N_r-1)/2} \qquad (2.3.15)$$

where:

SM_i = sum of all pairwise similarities among entities within cluster i.

N_i = number of entities in i.

♦ Average linkage within groups.

The average similarity for links between the two clusters is considered (Fig. 2.3.6e).

$$\frac{S_{tr}}{N_t N_r} \qquad (2.3.16)$$

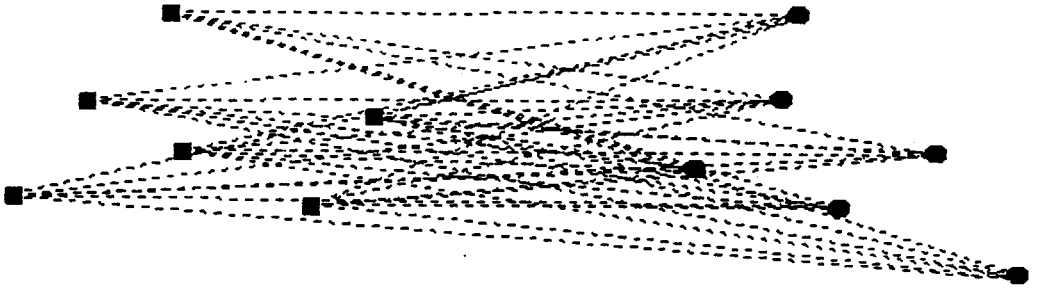

Figure 2.3.6e Average linkage within groups.

2.3.7 Dynamic Coalescence.

The dynamic coalescence model given by Watanabe (1985) forms clusters by merging points and clusters as in hierarchical clustering. The difference is that the points-clusters are actually moved in space by an attractive force toward one another. When two clusters touch, they are merged into one. The objective of dynamic coalescence is to minimize the entropy of the data. The dynamical motion of the points by mutual attraction will result in formation of more ordered structure (Fig. 2.3.7).

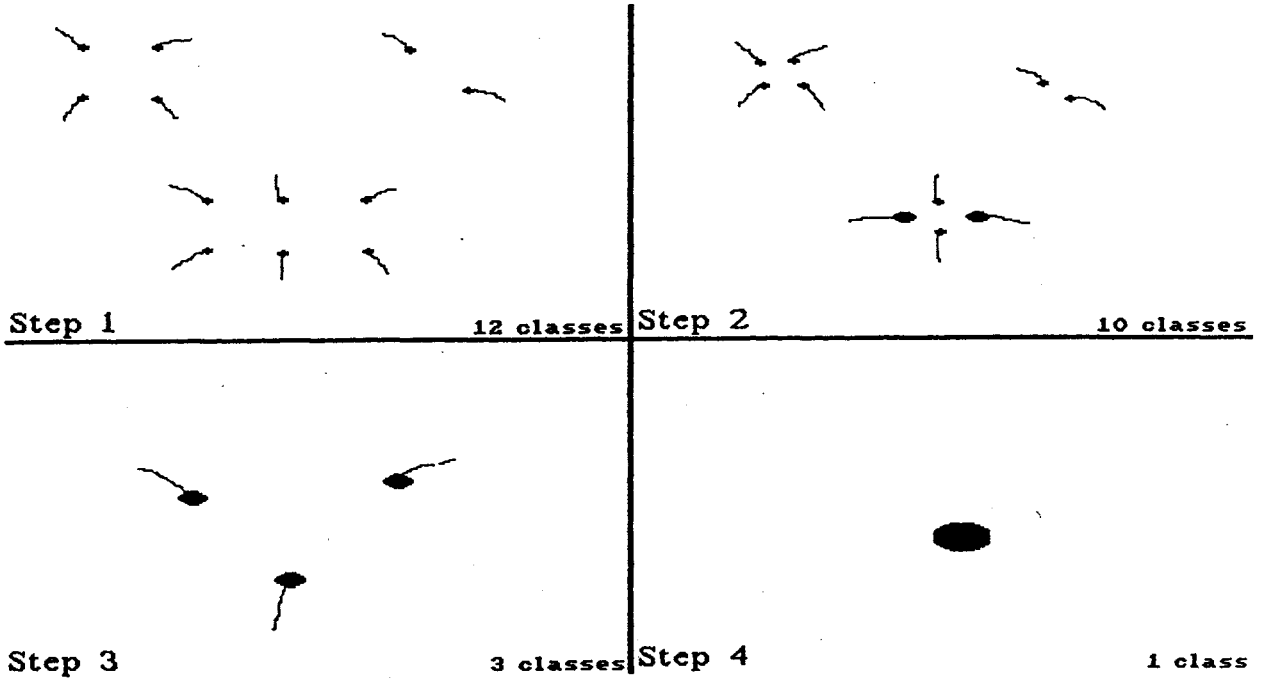

Figure 2.3.7 Illustration of dynamic coalescence algorithm.

If the position of cluster i is given by the coordinates of the vector z_i, the motion of this cluster is determined by the dynamic equation:

$$\frac{dz_i}{dt} = F_i \qquad (2.3.17)$$

where

$$F_i = A\sum_{j=1}^{N} \left[\frac{z_j - z_i}{|z_j - z_i|} (m_i m_j)^p \, g(|z_j - z_i|) \right] \qquad (2.3.18)$$

The effect of cluster j on i is a function of the mass, m, of i and j and the distance between i and j. A larger cluster will exert a greater force, and, therefore, p is greater than 0. The constant, A, is also greater than 0 since the force is attractive. The factor $g(\delta)$, where $\delta = | z_j - z_i |$, is a force function. Two types of force can be used:

- Gaussian:

$$g(\delta) = c \exp\left[-\frac{\delta^2}{2\sigma^2}\right] \qquad (2.3.19)$$

- Cauchy:

$$g(\delta) = c\frac{\sigma}{\delta^2 + \sigma^2} \qquad (2.3.20)$$

σ is a parameter that determines the strength of the force in relation to the distance between the two clusters. If i and j are further apart than σ, the force is very small. The algorithm is summarized below.

Step 1. Begin with each sample point, x_i, as a separate cluster, z_i.

Step 2. Calculate the change in position of each cluster for a small dt.

$$dz_i = F_i dt \qquad (2.3.21)$$

The new position of cluster i is then

$$z_i(t+dt) = z_i(t) + dz_i(t) \qquad (2.3.22)$$

Step 3. If the distance between two clusters is less than the sum of their radius, then the clusters are merged.

$$|z_i - z_j| \leq r_i + r_j \qquad (2.3.23)$$

The radius of a cluster is

$$r_i = (m_i)^{\frac{1}{n}} r_0 \qquad (2.3.24)$$

where r_o is the radius of each original sample point and n is the number of points in the cluster.

Step 4. The mass of the newly merged cluster is

$$m_i + m_j \qquad (2.3.25)$$

The position of the newly merged cluster is

$$\frac{z_i m_i + z_j m_j}{m_i + m_j} \qquad (2.3.26)$$

Step 5. Repeat until all clusters are merged into one.

Like hierarchical clustering, the outcome of the dynamic coalescence model does not depend on a starting point or points. The results will be affected by the values of the parameters A, p, and σ. Large values for these parameters cause large changes in position of the points for each dt. This can result in undesirable changes in the structure of the data.

2.4 Feature Selection and Extraction.

Before a pattern recognizer is designed, it is necessary to consider the feature extraction and data reduction problems. Any object or pattern which can be recognized and classified possesses a number of discriminatory properties or features. The first step in any recognition process, performed either by a machine or by a human, is to consider the problem of what discriminatory features to select and how to extract these

features. It is evident that the number of features needed to successfully perform a given recognition task depends on the discriminatory qualities of the chosen features. However, the problem of feature selection is usually complicated by the fact that the most important features are not necessarily easily measurable, or, in many cases, their measurement is inhibited by economic considerations.

It is evident that feature selection and extraction plays a central role in pattern recognition. In fact, the selection of an appropriate set of features which take into account the difficulties present in the extraction or selection process, and at the same time result an acceptable performance, is one of the most difficult tasks in the design of pattern recognition systems.

The feature extraction problem plays a central role in preprocessing and data reduction. This problem consists of determining certain invariant attributes of the pattern classes under consideration. These attributes are then used, for example, to reduce the dimensionality of the pattern vectors by means of a linear transformation. Once a set of attributes has been selected, the extraction process consists simply of extracting these attributes from the patterns under consideration (Fig. 2.3.8).

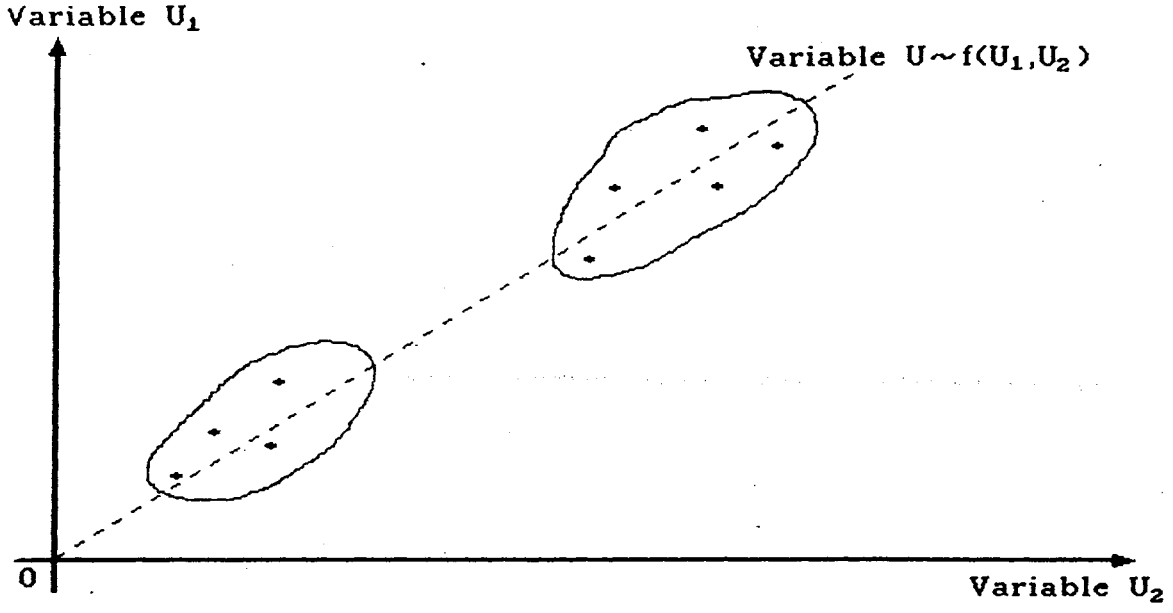

Figure 2.3.8 Simple example of feature extraction.

Pattern preprocessing generally involves two major tasks: clustering transformation and feature selection. A major problem in pattern recognition is the development of decision functions from sets of finite sample patterns of the classes so that the functions will partition the measurement space into regions each of which contains the sample pattern points belonging to one class. This argument leads to the concept of clustering transformation, which is made on the measurement space in order to cluster the points representing samples of the class. Such a transformation will maximize the intreset distance, while minimizing the intraset distance. The intreset distance is defined as the mean-square distance between pattern points that belong to two different classes.

The intraset distance is the mean-square distance between pattern points of the same class.

Through selection of the most effective features, the dimensionality of the measurement vector can be reduced. Feature selection may be accomplished independently of the performance of the classification scheme. Optimum feature selection is dictated by the maximization or minimization of a criterion function. Such an approach may be referred to as absolute feature selection. An alternative approach is performance-dependent feature selection , the effectiveness of which is directly related to the performance of the classification system, usually in terms of the probability of correct recognition. When the feature distribution for each pattern class in known, we may use divergence of entropy function in effecting feature selection. When the feature distribution for each pattern class is unknown, nonparametric feature selection based on direct estimation of the error probability may be used.

2.4.1 Feature Selection Through Entropy Minimization.

Entropy is a statistical measure of uncertainty. For a given ensemble of pattern vectors, a good measure of intraset

dispersion is the population entropy, given by:

$$H = -E_p\{\ln p\} \qquad (2.4.1)$$

where p is the probability density of the population, and E_p is the expectation operator with respect to p. The entropy concept can be used as a suitable criterion in the design of optimum feature selection. Features which reduce the uncertainty of a given situation are considered more informative than those which have the opposite effect. Thus, if one views entropy as a measure of uncertainty, a meaningful feature selection criterion is to choose the features which minimize the entropy of the pattern classes under consideration. Since this criterion is equivalent to minimizing the dispersion of the various pattern populations, it is reasonable to expect that the resulting procedure will have clustering properties.

Consider M pattern classes whose populations are governed by the probability densities $p(x/w_1)$, $p(x/w_2)$,..., $p(x/w_M)$. The entropy of the ith population of patterns is, from Eq. 2.4.1, given by:

$$H_i = -\int_x p\left(\frac{x}{w_i}\right) \ln p\left(\frac{x}{w_i}\right) dx \qquad (2.4.2)$$

where the integration is taken over the pattern space. It is observed that, if $p(x/x_i) = 1$, indicating no uncertainty, $H_i = 0$, in agreement with the previous interpretation of the entropy concept.

Assume that each of the M pattern populations is characterized by a normal probability density function, $p(x/w_i) <==> N(m_i, C_i)$, where m_i and C_i are the mean vector and covariance matrix, respectively, of the ith population. In addition, it will be assumed that the M covariance matrices describing the statistics of the M pattern classes are identical. With these assumptions in mind, the basic idea consists of determining a linear transformation matrix A, which operates on the pattern vectors to yield new vectors of lower dimensionality. This transformation may be written as:

$$\|y\| = \|A\| \times \|x\| \qquad (2.4.3)$$

where the transformation matrix is determined by minimizing the population entropies of the various pattern classes under consideration. In Eq. 2.4.3 x is a n-vector, y is an image m-vector of lower dimensionality than x, and A is an m x n matrix. The rows of the matrix A consist of the selected m feature vectors a_1', a_2',..., a_m', which are row vectors.

Thus, the matrix A is given by:

$$A = \begin{matrix} a_1' \\ a_2' \\ \vdots \\ a_m' \end{matrix} \qquad (2.4.4)$$

The problem is how to select the m feature vectors so that the measurement vector x is transformed to the image vector y while minimizing the entropy function defined by Eq. 2.4.2. A multivariate normal distribution is completely characterized by its mean vector and covariance matrix. This matrix is, in turn, characterized by its eigenvalues and eigenvectors. The eigenvectors may be regarded as the property vectors of the patterns under consideration. Some of the property vectors carry less information in the pattern recognition sense than others and may therefore be ignored. This phenomenon suggests a feature selection procedure whereby the most significant property vectors are chosen as feature vectors. These feature vectors can then be used to construct the transformation matrix A. After the formation of the matrix A, the number of vectors utilized should be large enough for the image patterns to carry sufficient discriminatory information.

2.4.2 Karhunen-Loève Expansion to Feature Selection.

The application of the K-L expansion to feature selection may

be viewed as a linear transformation. If we consider:

$$\Omega = (\Omega_1, \Omega_2, \ldots, \Omega_m), \; m<n \tag{2.4.5}$$

to be the transformation matrix, then, from Eq. 2.4.5, the image patterns are the coefficients of the K-L expansion, that is, for any pattern x_i of class w_i we know:

$$e_i = \Omega' x_i \tag{2.4.6}$$

Since Ω' is an m x n matrix and x is an n-vector, we see that, if $m < n$ the e_i are image vectors of lower dimensionality.

It can be shown that the optimum properties of K-L expansion are satisfied if the columns of the transformation matrix Ω are chosen as the m normalized eigenvectors corresponding to the largest eigenvalues of the correlation matrix R. The above notation can be expressed by defining the matrix:

$$A = \Omega' = \begin{matrix} \Omega_1' \\ \Omega_2' \\ \vdots \\ \Omega_m' \end{matrix} \tag{2.4.7}$$

where the rows of A are now the normalized eigenvectors corresponding to the largest eigenvalues of R. If we let y = e, then, for any vector x, the reduced image vectors are given

by:

$$[y]=[A]\times[x] \qquad (2.4.8)$$

the foregoing results may be summarized as follows:

- ♦ Compute the correlation matrix R from the patterns of the training set.
- ♦ Obtain the eigenvalues and corresponding eigenvectors of R. Normalize the eigenvectors.
- ♦ Form the transformation matrix Ω from the m eigenvectors corresponding to the largest eigenvalues of R.
- ♦ Compute the coefficients of the expansion. These coefficients represent the reduced image patterns.

Although the assumption that all pattern populations must have identical means is certainly a limitation of the K-L expansion, one should not conclude that this approach to feature selection is without merit. Assumptions such as this one are characteristic of most statistical methods of analysis. The success of any given method depends simply on how closely the data under consideration conform to the basic assumptions underlying the development of the statistical technique.

2.5 Classification Success Index.

In most real life clustering situations, a researcher is faced with the dilemma of selecting the number of clusters or partitions in the final solution (Everitt, 1979; Sneath and Sokal, 1973). Virtually all clustering procedures provide little if any information as to the number present in the data. Nonhierarchical procedures usually require the user to specify this parameter before any clustering is accomplished and hierarchical methods routinely produce a series of solutions ranging from n clusters to a solution with only one cluster present (assume n objects in the data set). As such, numerous procedures for determining the number of clusters in a data set have been proposed (Dubes and Jain, 1979; Milligan, 1981). When applied to the results of hierarchical clustering methods, these techniques are sometimes referred to as stopping rules. Often, such rules can be extended for use with nonhierarchical procedures as well.

The application of a stopping rule in a cluster analytic situation can result in a correct decision or in a decision error. Basically, two different types of decision errors can result. The first kind of error occurs when the stopping rule indicates k clusters are present when, in fact, there were less than k clusters in the data. That is, a solution

containing too many clusters was obtained. The second kind of error occurs when the stopping rule indicates fewer clusters in the data than are actually present. Hence, a solution with too few clusters was obtained. Although the severity of the two types of errors would change depending on the context of the problem, the second type of error might be considered more serious in most applied analyses because information is lost by merging distinct clusters.

In this study the Calinski and Harabasz (1974) index was used to determine the number of clusters. This index is computed as:

$$\frac{\frac{trace\ B}{k-1}}{\frac{trace\ W}{n-k}} \qquad (2.5.1)$$

where n and k are the total number of items and the number of clusters in the solution, respectively. The B and W terms are the between and pooled within cluster sum of squares and cross products matrices. The maximum of this index indicates the correct number of partitions in the data.

DATA DESCRIPTION

3 Data Description.

Many conference sessions, workshops and professional papers have been devoted to the merits of different methods of performance analysis and the merits of specific performance measures. Miller (1980) has summarized this literature and suggested the need for a simple framework using three or four measures of efficiency and effectiveness. In this paper a set of indicators are used to classify the transit systems. And further a small unique set of indicators can be selected from these concepts to give the same classification results.

For this study the three following sets of data were used:

- Operating characteristics of the systems.
- Socioeconomic characteristics of the service area.
- Ratios formed by the variables in the other two sets.

Some data was obtained from the Department of Transportation which collects this data from all transit systems that request financial assistance from MN/DOT. Also data was collected from other sources, such as, Department of Revenue, and Minnesota State Demographer.

3.1 Systems Collected.

All transit systems in the state of Minnesota that apply for assistance from the state are included in the analysis. MN/DOT currently uses a peer grouping based on population and type of service provided. The eight peer groups are listed in the following Table 3.1.1. One goal of this clustering analysis will be to compare the results with the peer groups established by MN/DOT. Some problems can be found in the present peer grouping. Some small urban systems perform more like some of the urban systems. Also, there is a problem with the population criterion for small urban and rural systems. The criterion is used when the transit service is within a city, but not for counties. Although most counties have populations over 2,500, they are considered rural because they operate in rural areas outside of a city.

3.2 Variables.

Not all researchers agree that multiple measures of transit performance are needed. Several authors have advanced claims that a single measure is sufficient. Nash (1980) prefers cost per passenger or passenger per mile when analyzing alternative investments for management. Patton (1983) has suggested that

Table 3.1.1 Peer groups proposed by Mn/DOT.

1. Metropolitan Transit Commission
   1 System
   Fixed route
   Minneapolis - St. Paul metropolitan area

2. Large Urban
   1 System
   Fixed route
   Duluth Transit Authority

3. Twin Cities Regular Route
   1 System
   Fixed Route
   Private Operators
   Minneapolis - St. Paul suburbs

4. Twin Cities Dial-A-Ride (Metro Mobility)
   3 Systems
   Specialized service
   Minneapolis - St. Paul metropolitan area

5. Urban
   3 Systems
   Fixed route
   Population over 50,000

6. Urban Dial-A-Ride
   4 Systems
   Specialized service
   Population over 50,000

7. Small Urban
   24 Systems
   Fixed route and/or Specialized service
   Population 2,500 to 50,000

8. Rural
   19 Systems
   Fixed route and/or Specialized service
   Population under 2,500

transportation statistics on performance can be integrated about the indicator of cost per passenger. Each assumes that the overreaching goal for transit is transporting passengers for the minimum cost. Kneafsey (1975) refers to this as "efficiency-in-the-small" or cost minimization to the firm. Allocative efficiency, what Kneafsey calls "efficiency-in-the-large", is more suitable for transit performance analysis. This defines efficiency in terms of resources used to produce service. By this definition, efficiency is a statement about the achievements of an agency in transforming a set of inputs into a set of outputs. Others have used "technological efficiency" for this concept in public sector analysis in contrast to "economic efficiency".

In order to create a valid and applicable classification of bus transit systems, several considerations were used to select an appropriate data base. The source of data should be reliable with all the public transit systems in Minnesota represented. The variables from the data base should be comparable across systems. They should be compiled using standard definitions and in similar ways for each agency and they should be validated for accuracy. The variables used to form the data base should be verified from other studies that they are indeed related to transit performance and they should also be easily understood and used by the transit community.

Several factors affect the bus transit performance either directly or indirectly. As such these factors are potentially useful in establishing a classification scheme that can be used to explain the variation among bus operations. As the data is the only criterion to classify the transit systems, should be developed to meet some requirements. The data must reflect the main characteristics of a transit system and the changes occurring in a transit system over time. Also the data has to be available for every transit system and reliable (Table 3.2.1).

Confusion over defining efficiency has reduced the value of many studies. Public transit agencies can not focus on a single objective function as they must respond to the objectives of various "publics". Advocacy of single measures of transit performance integrating efficiency and effectiveness has not aided transit performance. Separate measures provide more useful results.

From a large set of possible variables a set of data was chosen that could reflect the service area characteristics that constrain the decisions made by transit operators. For each transit system, three different sets of data have been introduced. The first is composed of operating characteristics of the transit systems, the second is a set of

Table 3.2.1 Summary of the collected data.

INDICATORS	ESTI-MATED	COUNTY CITY TRANSIT	SOURCE OF DATA
Transportation data			
Operating Cost	NO	TRANSIT	MN / DOT
Operating Revenue	NO	TRANSIT	MN / DOT
Government Subsidy	NO	TRANSIT	MN / DOT
Vehicle Miles	NO	TRANSIT	MN / DOT
Vehicle Hours	NO	TRANSIT	MN / DOT
Passengers	NO	TRANSIT	MIN. TRANSIT REPORT
Vehicles	NO	TRANSIT	MIN. TRANSIT REPORT
Employees	NO	TRANSIT	MIN. TRANSIT REPORT
Socioeconomic data			
Population	YES	CITY	MIN. STATE DEMOGRAPHER
Income	NO	COUNTY	MIN. STATE DEMOGRAPHER
Households	YES	CITY	MIN. STATE DEMOGRAPHER
Driver's License	YES	CITY	MOTOR VEH. CRASH DATA
Unemployment	YES	COUNTY	MINNESOTA IN THE 80's
Poverty Level	YES	COUNTY	MINNESOTA IN THE 80's
Sales Tax	NO	COUNTY	DEPARTMENT OF REVENUE
Property Tax	NO	COUNTY	DEPARTMENT OF REVENUE
Population Density	YES	COUNTY	MIN. STATE DEMOGRAPHER
Average Age	NO	COUNTY	MIN. STATE DEMOGRAPHER
Hi/way Expenditures	NO	COUNTY	DEPARTMENT OF REVENUE

socioeconomic indicators of the area served by the system, and the third set consist of ratios formed by the variables in the other two sets (Fig. 3.2.1).

From the first set, eight aspects of transit systems were used to characterize transit systems variables, employees, cost, revenue, government subsidy, vehicle hours, vehicle miles and

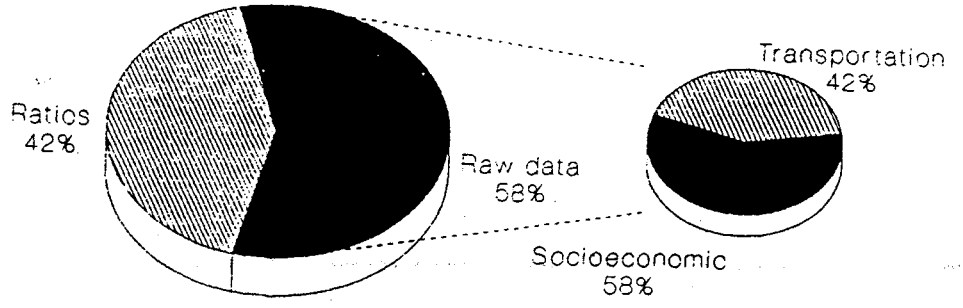

Figure 3.2.1 Type of data used.

passengers. The number of vehicles and the number of employees indicate the size of the system. Size reflects a number of constraints of transit management. Organized labor units are more influential in larger agencies. Efficient route scheduling is more difficult, and managing large numbers of employees is more complex. These factors cause diseconomics of scale reducing the advantages gained through service integration. Very small systems also suffer from constraints that restrict efficient use of resources. Operating cost include all costs of the transit system, wages, capital expenses, maintenance and repair, contracts for private operators, insurance and taxes. Revenue consists primary of fare box revenues but also include special

contracts with institutions, such as Universities and large businesses along with income from charter services. Government subsidy is the total assistance provided by local state and federal governments, to make up for the difference in transit systems, operating cost and operating revenue.

Total vehicle miles and total vehicle hours are the annual miles and hours compiled by the system's fleet. Vehicles include the number of operating and back-up vehicles. Operating and government subsidy, vehicles, and employees are known as service inputs. They represent resources which are used by the system. Total vehicle miles and total vehicle hours are service outputs, and they show what the system provides to the service area. Operating revenue and passengers are measures of service consumption. The number of passengers indicates how much the service is used, and operating revenue similarly reflects use of the service since revenue increases as use increases.

The second set of data has eleven socioeconomic factors of the service area. Population, households, driver's license, income, sales tax, property tax, highway expenditures, population density, age, unemployment, poverty level. The service area population is a socioeconomic factor which may indicate the potential market for the system. If an indicator

of productivity such as passengers per revenue vehicle mile, for example, is used as a measure of transit performance, consideration must then be given to the population characteristics of the community, the system serves. If level of service are equal, passenger use has been found to vary directly with population (Sinha C. K., 1980). Accordingly, total population of the urban area is the variable used in the system classification. Number of households and number of driver's license give the need for transit in the area.

The socioeconomic status of the area is given by the income, property tax, highway expenditures and sales tax. Poverty level and unemployment can be used to determine the type of service, the system has to provide. These measures reflect the need to serve special groups such as the elderly, handicapped and people with low income. Failure to recognize urban areas that have large concentrations of such special population groups may result in inaccurate assessments of transit performance. In general, it would seem reasonable to assume that those transit systems that provide a high level of service to the elderly and the handicapped would occur higher unit operating costs than do those systems in cities with a small elderly and handicapped population.

There is a little disagreement that the population as well as

the area of a city play some role in both the provision and the consumption of transit service. Although size of the population alone can give some indication for the ridership levels, additional insight into transit performance can be gained if land area is integrated into the analysis. In general, transit service is more efficiently an defectively provided in high density areas. Furthermore, transit operational and financial performance is affected not only by density of residential population but also by the density and size of nonresidential (i.e. industrial and commercial) clusters in an urban area. The significance of such a relationship permits estimation of the effect of different land-use policies on transit performance.

The third set of the data are ratios of the other two data set. They are known as performance indicators since they are often used to determine the performance of the transit system. There are three primary types of performance indicators, cost efficiency, cost effectiveness, and service effectiveness. Cost efficiency indicators measure the service inputs (labor, capital, fuel) to the amount of service produced (service outputs: vehicle hours, vehicle miles). Cost effectiveness indicators measure the level of service consumption (passengers, operation revenue) against service inputs. Finally service effectiveness indicators measure the extent to

which service outputs are consumed.

3.3 Missing Data.

Missing information poses a unique analytical problem. Both valid zeros and "no information reported" codes are represented by zeros (Fig. 3.3.1). Whenever possible, other information available in the data base was pieced together to provide for missing data or to distinguish between valid zeros and failure to report. Missing values encountered at any point in the computation of basic and ratio variables and during statistical procedures cause problems and make the results insignificant. The missing values problem has a cumulative effect as factors are dropped from the analysis.

3.4 Data Analysis.

One objective of this research is to establish a small, unique subset of indicators that is particularly useful for classifying the transit systems. The goal is to identify the minimum amount of data necessary to convey the maximum amount of information. The dimensionality of the data is reduced in two ways, feature selection and feature extraction.

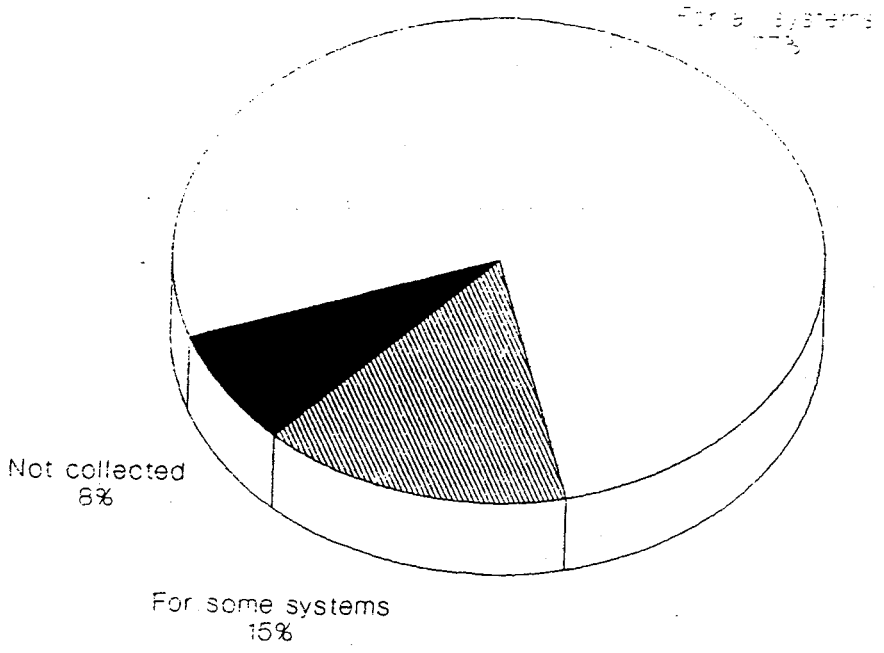

Figure 3.3.1 Collected data.

Feature selection was used to remove highly correlated variables. When two or more variables were highly correlated only one was not removed from the set. To select this variable the following criteria were used:

- Representativeness of the difference between the systems by this variable.
- The distribution of values in the variable has to be as close to normal as possible.
- Ease of collection of the variable was assessed by the percentage of data missing.
- The variable selected has to be easily understood by transit managers.

Feature extraction is the mapping of a set of points in N-dimensional space to M-dimensional space where M is less than N. This is done because some measurements are redundant or highly correlated. The dimensionality can thus be reduced with little loss of information. Also computations, for classification are easier when the dimensionality is reduced. Furthermore in the case where the sample is reduced to two or three dimensions, the data is physically more meaningful and can be plotted.

Karhunen-Loève is a linear feature extractor mapping $\|x\|$ to $\|y\|$ via a q by p matrix $\|T\|$ such that $\|y\| = \|T\| \|x\|$. In the Karhounen-Loève expansion, T is composed of the eigenvectors corresponding to the q largest eigenvalues of the covariance matrix of x. Before Karhounen-Loève was applied, the both data sets, operating and socioeconomic, were normalized. This was done so that variables with large magnitudes would not receive more weight in the clustering analysis.

METHODOLOGY

4 Methodology.

4.1 Classification Algorithms Applied.

In this study two methods were used to classify the transit systems, the combined method and the new method. The combined method is a combination of k-means and maximum distance classification methods. This method was used because it is fast, easy to implement and gives fairly good results. On the other hand the new method gives much better results and can handle more complicated data structures but it needs more computation time.

4.1.1 Combined Method.

This method is not actually a new method, is a combination of two existing classification methods, k-means and maximum distance. These two algorithms were chosen for further examination because of their satisfactory results.

The systems chosen by the maximum distance algorithm as cluster centers were used as the initial cluster centers for the k-means algorithm. This improved both algorithms because:

- The initial points for k-means are now more likely to be spread evenly throughout the data.

♦ Instead of just assigning the remaining points to the nearest cluster center found by the maximum distance, these cluster centers are only used as initial cluster centers for the k-means algorithm.

Also in order to eliminate some disadvantages, the two algorithms were used with some changes. The results now are better than by using the k-means or the maximum distance algorithm alone.

Suppose that we have a set of N sample patterns $\{x_1, x_2, \ldots, x_N\}$. This procedure requires only the desired number of clusters K. The two parts of the algorithm are described by the following steps.

Part 1. (maximum distance).

Step 1. For all the pairs of points x_i and x_j compute the distances d_{ij} between these points and save the maximum of these distances.

$$\max d\text{-}\max\{d_{ij}[x_i, x_j],\ \forall i, j,\ i \neq j\} \qquad (4.1.1)$$

Step 2. The two points x_a, x_b with the maximum distance d_{ab} become the first two cluster centers z_1 and z_2.

Step 3. For z_3 to z_K:

Step 4. Compute the distance from each remaining

sample, x_i, to all z's. Save the minimum of these distances for each x_i.

$$MIN_i = \min\{\|x_i - z_j\|, \ \forall j\} \qquad (4.1.2)$$

Step 5. Consider the sample with the largest minimum distance.

$$MAX = \max\{MIN_i, \ \forall j\} \qquad (4.1.3)$$

If the z_K cluster center was calculated then go to step 6 else go to step 4.

Part 2. (k-means).

Step 6. Set the k initial cluster centers $z_1(1) = z_1$, $z_2(2) = z_2, \ldots, z_K(1) = z_K$.

Step 7. At the kth iterative step distribute the samples {x} among the K cluster domains, using the relation:

$$x \epsilon S_j(k) \ \text{ if } \|x - z_j(k)\| < \|x - z_i(k)\|, \quad \forall i = 1,2,\ldots,K, \ i \neq j \qquad (4.1.4)$$

where $S_j(k)$ denotes the set of samples whose cluster center is $z_j(k)$.

Step 8. From the results of step 7, compute the new cluster centers $z_j(k+1)$, j = 1, 2,..., K, such that the sum of the squared distances from all points in $S_j(k)$ to the new cluster center is minimized. In other words, the new cluster center $z_j(k+1)$ is

computed so that the performance index:

$$J_i = \sum_{x \in S_j(k)} \|x - z_j(k+1)\|^2, \quad j = 1, 2, \ldots, K \qquad (4.1.5)$$

is minimized. The $z_j(k+1)$ which minimizes this performance index is simple the sample mean of $S_j(k)$. Therefore, the new cluster center is given by:

$$z_j(k+1) = \frac{1}{N_j} \sum_{x \in S_j(k)} x, \quad j = 1, 2, \ldots, K \qquad (4.1.6)$$

where N_j is the number of samples in $S_j(k)$.

Step 9. If $z_j(k+1) = z_j(k)$ for j = 1, 2, ..., K, the algorithm has converged and the procedure is terminated. Otherwise go to step 7.

The results of the combined method does not depend on any initial selections or any threshold value. In all the cases that the combined method was tested gave better or the same results than k-means or maximum distance methods (Fig. 4.1.1).

4.1.2 New Method.

Cluster seeking may be viewed as a problem in unsupervised pattern recognition. Suppose that we are given a set of patterns without any information whatsoever as to the number of classes present in the group. The unsupervised learning problem may be stated as that of identifying the classes in

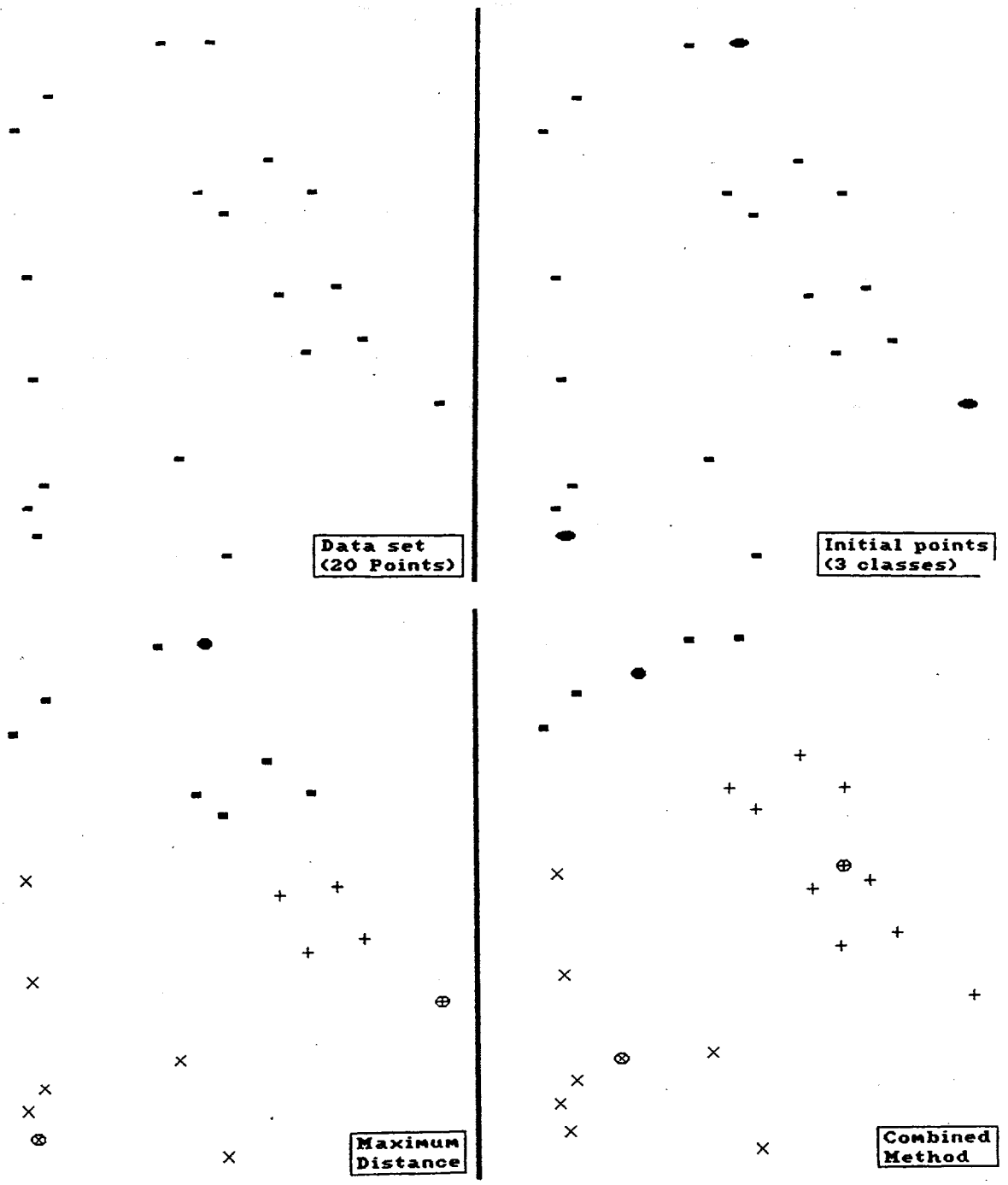

Figure 4.1.1 Application of combined and maximum distance methods on the same data set.

the given set of patterns. If we are willing to accept cluster centers as a method of representation, one obvious way of characterizing a given set of data is by cluster identification. The application of cluster-seeking algorithms to unsupervised learning is, in principle, straightforward. Suppose that we are given a set of patterns $\{x_1, x_2, \ldots, x_N\}$ of unknown classification. These patterns may be submitted to one or more algorithms in an effort to identify representative cluster centers. The resulting cluster domains may then be interpreted as different pattern classes.

Being mechanical simulations of human perception, the methods of pattern recognition may profit from imitating a certain special aspects of human perception. In trying to see a "form" in a collection of points, the formation of patterns will be aided by an imaginary connection among the points. The areas of the data space were the points are "dense" are considered to include cluster centers (Fig. 4.1.2). But a classification method needs decision functions which generate the partition boundaries in the pattern space to separate patterns of one class from another. The method, that is described, is using the potential function concept to determine the decision functions and the partition boundaries.

Suppose that we want to distinguish between two pattern

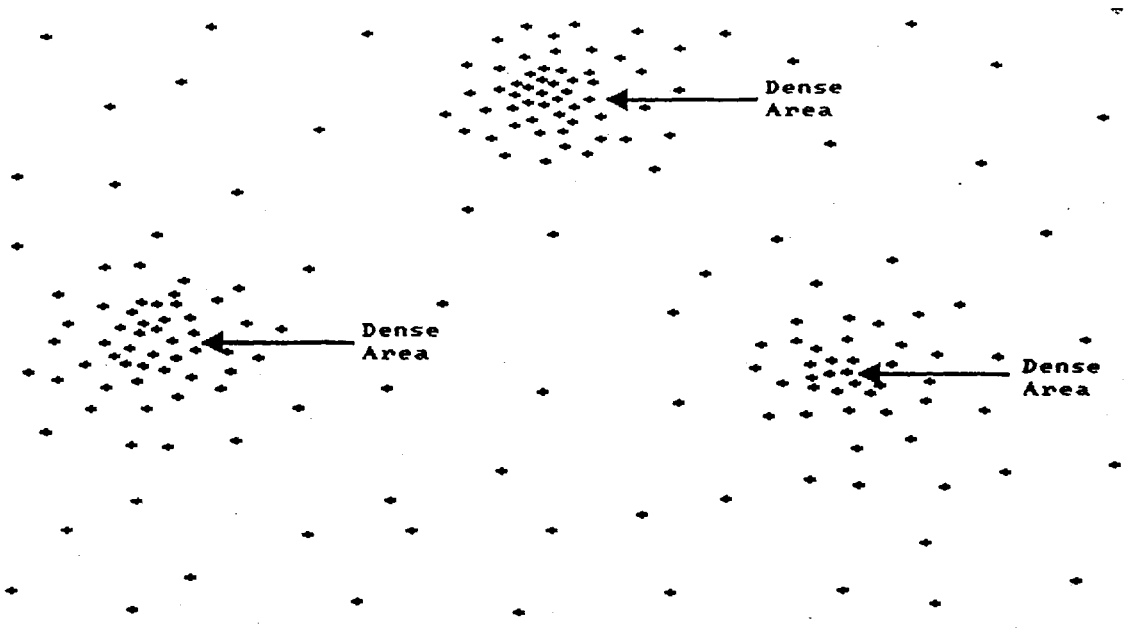

Figure 4.1.2 Dense areas of points in a data set.

classes, w_1 and w_2. Sample patterns of both classes are represented by vectors or points in the n-dimensional pattern space. If these sample pattern points are likened to some kind of energy source, the potential at any of these points attains a peak value and then decreases rapidly at any point away from the sample pattern point, x_k (Fig. 4.1.3). Using this analogy, we may visualize the presence of equipotential contours which are described by a potential function $K(x, x_k)$. For pattern class w_1, we may imagine that the cluster of sample patterns forms a "plateau" with the sample points located at the peaks of a group of hills. A similar geometrical interpretation may be visualized for pattern class w_2. These two "plateaus" are separated by a "valley" in which

the potential is said to drop to zero (Fig. 4.1.4).

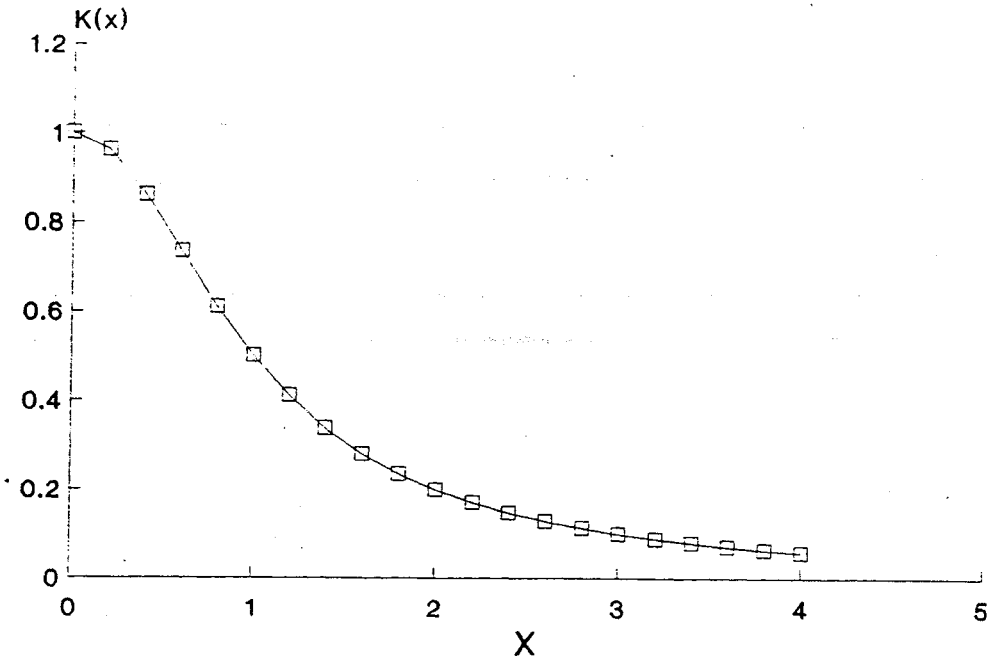

Figure 4.1.3 Potential function for one point.

Although an infinite series expansion is often employed in mathematical discussions of potential function algorithms, it clearly is of no practical usefulness. Usually a symmetrical function of two variables, x and x_k, is used as a potential function. By using symmetrical function we can see that $K(x, x_k) = K(x_k, x)$. Functions which can be used as potential functions are :

$$K(x, x_k) = \exp\{-a\|x - x_k\|^2\} \qquad (4.1.7)$$

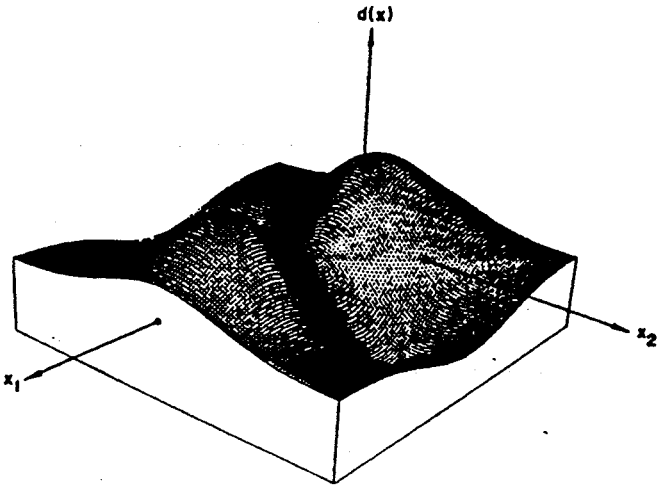

Figure 4.1.4 Case of two classes representation.

$$K(x, x_k) = \frac{1}{1+a\|x-x_k\|^2} \qquad (4.1.8)$$

$$K(x, x_k) = a\left(\frac{1-\|x-x_k\|^2}{a^2}\right), \quad x \le a$$
$$else\ K(x, x_k) = 0 \qquad (4.1.9)$$

where a is a positive constant, and $\| x - x_k \|$ is the norm of the vector $(x - x_k)$. It is worth noting that these functions are inversely proportional to the squared distance measure, $D^2 = \| x - x_k \|^2$, which is also a characteristic, for example, of the force in a gravitational potential field. The above functions are plotted in Fig. 4.1.5 for one-dimensional patterns and in Fig. 4.1.6 for two-dimensional case.

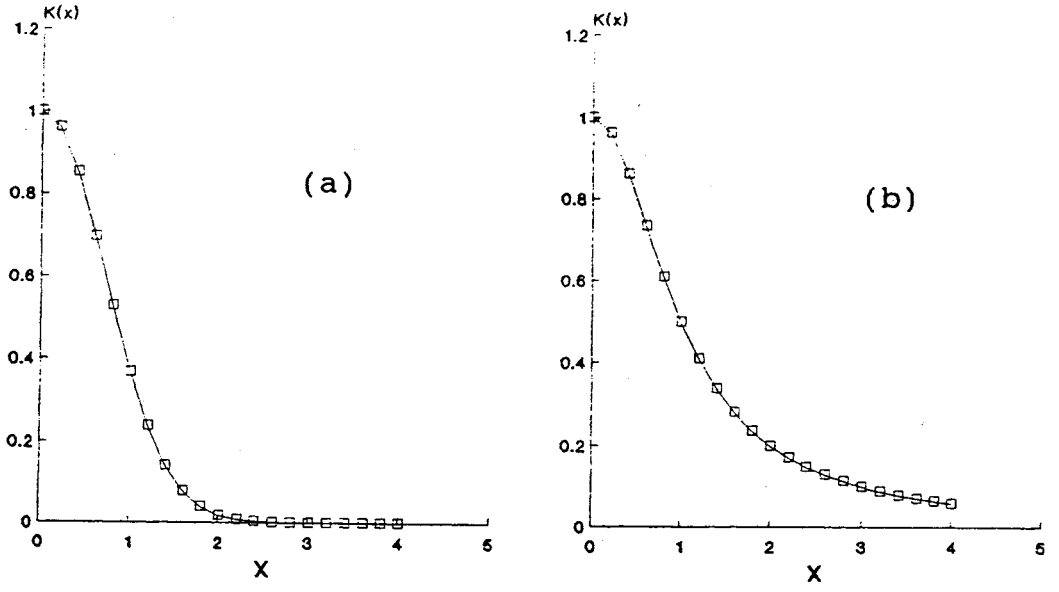

Figure 4.1.5 Examples of one-dimensional potential functions (a - Eq. 4.1.7, b - Eq. 4.1.8).

The value of a in Eq. 4.1.7, 4.1.8, 4.1.9 determines the area that each point "affects" other points. When this area is reasonably large the resulting number of classes is small and when the area is small the resulting number of classes is large. Before the application of the method either the desired number of classes or the value of a should be defined, the method will determine then the value of a or the number of classes respectively. In this study, always, the number of classes was given and the value of a was determined by the method. The basic steps of the algorithm are summarized below.

Step 1. Choose the number of classes K.

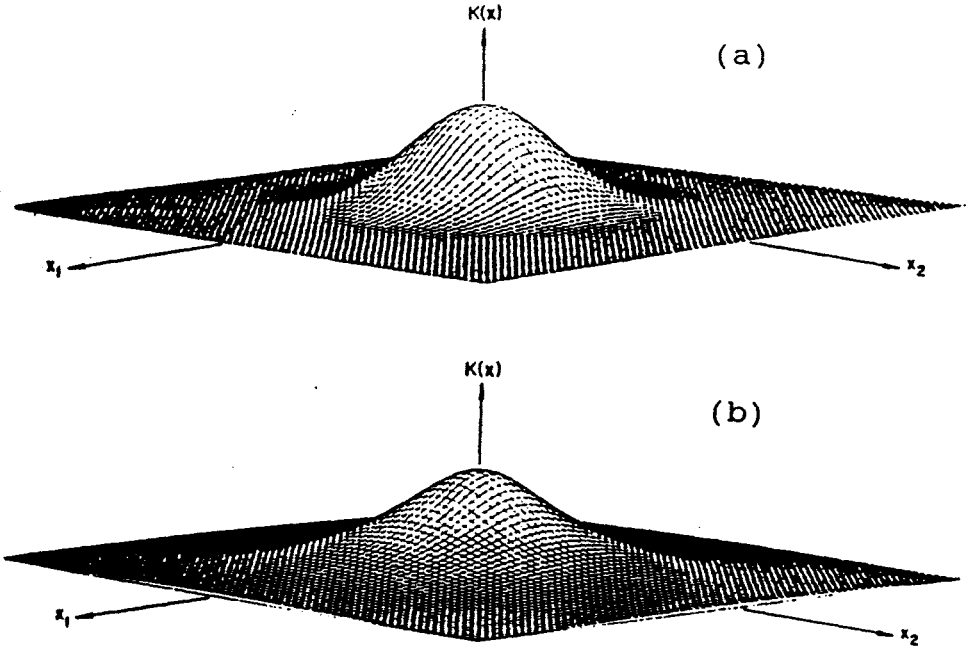

Figure 4.1.6 Examples of two-dimensional potential functions (a - Eq. 4.1.7, b - Eq. 4.1.8).

Step 2. Make a choice for the value of a.

Step 3. Apply the potential function to every data point and determine the cumulative function from all the points.

Step 4. Determine the resulting number of classes K_a.

Step 5. If $K < K_a$ or $K > K_a$ then decrease or increase, respectively, the value of a and go to step 3.

In step 4 the method is using a procedure to determine the number of classes. This procedure will be presented by two

examples.

Example 1. Let us apply the method to the patterns shown on Fig. 4.1.7a using the potential function of Eq. 4.1.9. Consider that we have cost per mile data for three transit systems A, B, and C and we want to have two classes. First the cumulative potential function is calculated. Then all the pairs of data points are determined. For each pair P_i and P_j the line segment between the two data points is defined and the value of the cumulative potential function is compared with a threshold T, for all the points on the line segment.

$$Compare\ T, K(x)\ \forall x \in [P_i, P_j] \qquad (4.1.10)$$

If at any point the value of the cumulative potential function K(x) is less or equal with the threshold T then the two data points are assigned in different classes, otherwise the two data points are assigned in the same class. For the first pair of points (Fig. 4.1.7b), A-B, the cumulative potential function is lower than the threshold T between points x_1 and x_2, so the systems A and B do not belong to the same class. The same result we have for the systems A-C. For the last pair B-C the function never goes under the threshold T and the two systems are assigned in the same class. Considering the results from all the pairs, we have one class with the system A and a second class with systems B and C, which is the

correct result.

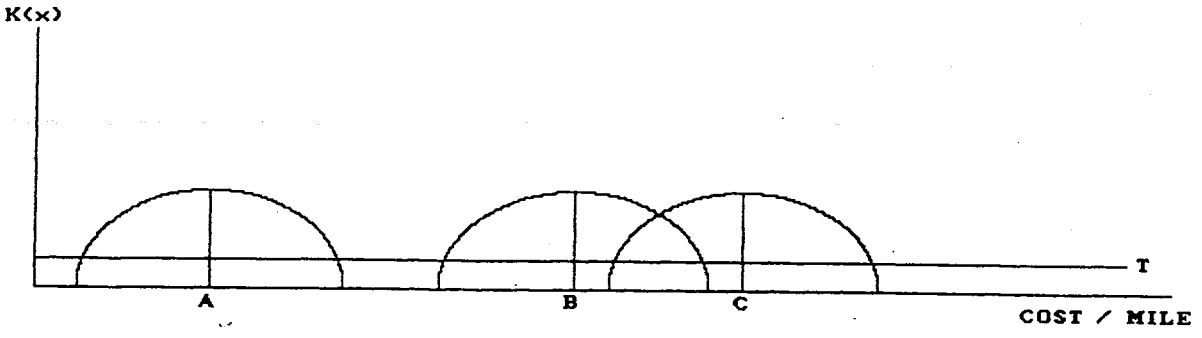

Figure 4.1.7a Data patterns used in example 1.

Example 2. When for each system there is information for more than one indicators then the patterns are more complicated. Consider the case that we have a data set with five transit systems and we have information for cost per mile and the number of passengers. On Fig. 4.1.8a we can see the data set and the potential function on each data point. After calculating the cumulative potential function we can have a top-view of the data (Fig. 4.1.8b). Now there are ten possible pairs. The value of the threshold T depends on the desired number of classes K. In the case that K = 2 then T = T_1. On Fig. 4.1.8c there is a side-view of the cumulative potential function for the pairs B-C and C-D. Using the same

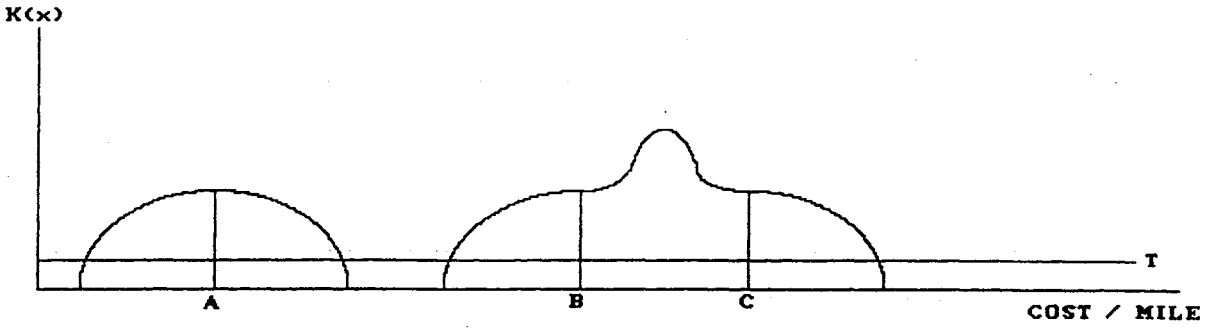

Figure 4.1.7b Cumulative potential function for example 1.

procedure as in the example 1 we have that systems B and C belong to the same class and system D belongs to a different class. The final result is that systems A, B and C belong to the first class and systems D and E belong to the second class. If the desired number of classes is K = 3 then T = T_2. By looking the Fig. 4.1.8c and examine the two pairs of data points B-C and C-D, we have that the three systems belong to three different classes. The final result is that systems A and B belong to the first class, the system C belongs to the second class, and the systems D and E belong to the third class.

In cases with more than two indicators it is not possible to

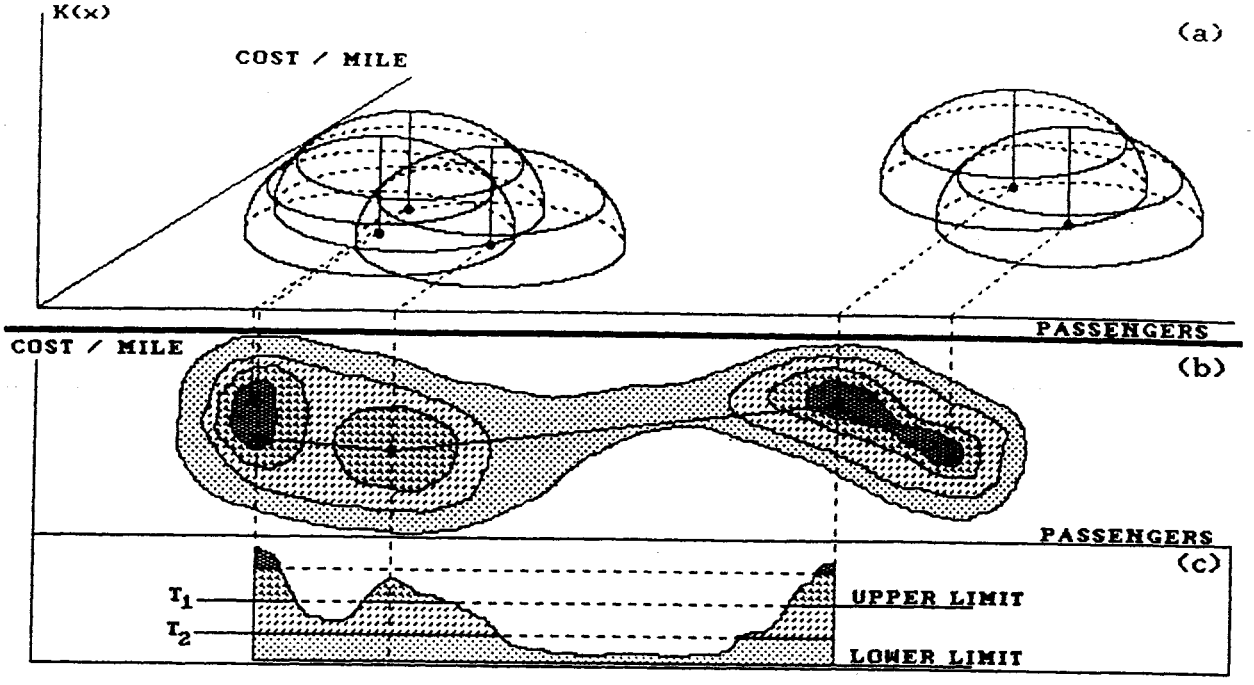

Figure 4.1.8 a. Data patterns used in example 2.
b. Top-view of the cumulative potential function.
c. Side-view of the cumulative potential function.

have a view of the data, but the method can be applied and give satisfactory results. This method has the following advantages:

♦ Because each point is under the influence of all other points, by using the appropriate potential function, the model is capable of reflecting the global situation of the distribution of all points. This is a great advantage over many other algorithms which consider only the influence of neighboring points.

- Instead of searching the optimum or local optimum by methods such as hill-climbing, exchange-of-members, and trial-and-error, the suggested algorithm seeks directly a global solution for the data set.
- The algorithm takes into account the chain effect in the data.
- We can obtain any number of clusters, ranging from one to N_o (the number of the initially given objects).

In the two following figures there are illustrations of the method. Fig. 4.1.9 shows that the method is suitable to the case where the chain effect is marked. In this example some of the two points within the same group are separated by a distance larger than the distance between some of the two points belonging to two different groups. The algorithm embodies such a "chain effect". In Fig. 4.1.10 the configuration of the points is produced by two normal distributions. This example shows that the method is very sensitive to the global situation of the distribution of points.

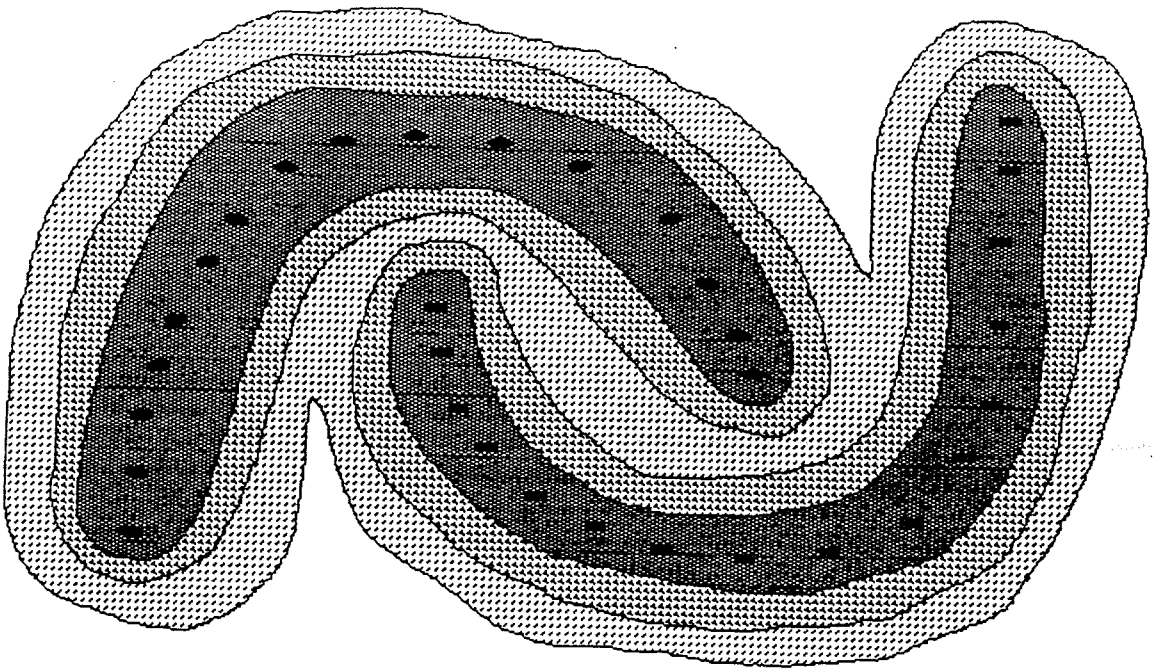

Figure 4.1.9 Data set where the chain effect occurs.

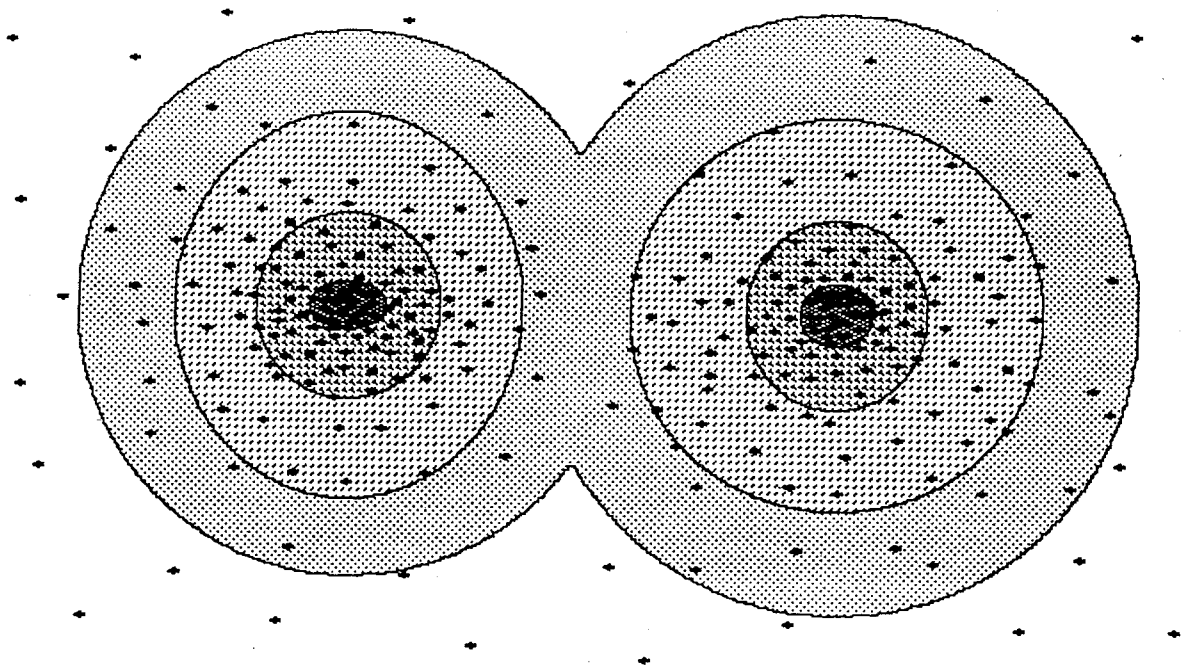

Figure 4.1.10 Data set produced by two normal distributions.

4.2 Methodology Followed.

In this section, the procedure that will be used in section 4.3, will be described. First, data reduction and data refinement is introduced. It is often desirable to reduce the number of variables used in a clustering analysis. One reason for reducing the dimensionality is that there may be some redundancy in the variables, i.e. some variables may be highly correlated. In such cases the dimensionality can often be reduced with little lose of information. Also, computations are simplified when there are fewer variables.

Before the classification methods were applied, the data set was scaled. This was done so that variables with large magnitudes would not receive more weight in the clustering analysis. The data set had a wide range of variation in the variables so it appeared appropriate to scale or to normalized the data in the set. When the data were scaled a number was selected, for each variable, so all the data will fall in a specified range. By using this method all the variables will be considered with the same weight in the clustering analysis. On Fig. 4.2.1 there is an example of two dimensional data set. The data were scaled in a range [0, 100]. In the case that the data were normalized the average and the standard deviation of each variable was calculated. Then from each

case the average value was subtracted and the result was divided by the standard deviation to form a new variable that will be used for the clustering analysis. After the normalization all the variables have average equal to zero and standard deviation equal to 1. After the normalization all the variables will be considered with the same weight in the clustering analysis. On Fig. 4.2.2 the data set that was used for the example on Fig. 4.2.1 is normalized.

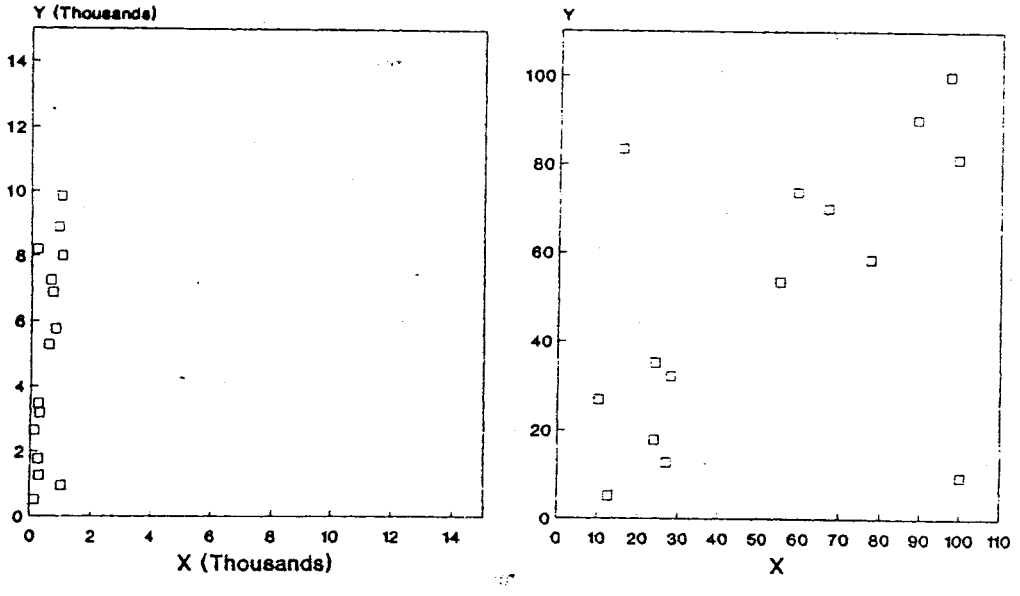

Figure 4.2.1 Example of a data set that was scaled.

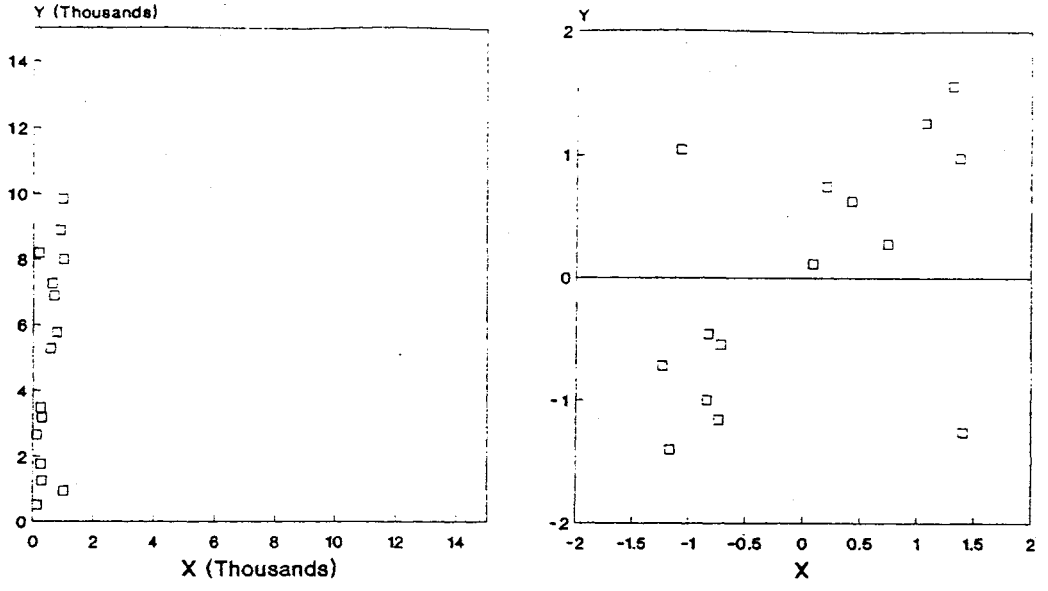

Figure 4.2.2 Example of a data set that was normalized.

After the data reduction and the data refinement was introduced the classification methods were applied to the data set. Because of the better performance the new method was selected for the clustering analysis. Different ways to classify the transit systems were tested, as it will be described in chapter 5, in order to get a meaningful classification. The results of the classification method were used as prior information to perform Karhunen-Loève expansion. After the feature selection we can have a data set, smaller than the data set used in the initial classification, with a little loss of information. If we repeat the cluster analysis with the new data set the result should be the same or close to the initial result.

In order to test how the method will classify new information, e.g. an new transit system, the data set was divided in two data sets a sample set and a test set. The data refinement and the clustering analysis were repeated by using only the sample data set. By using the classification results as prior information the Karhunen-Loève expansion method was used to obtain the most significant features. Then, by using these features the test data set was classified and a percent of misclassifications was calculated.

ANALYSIS OF THE RESULTS

5 Analysis of the Results.

5.1 Description of Different Classification Approaches.

In this section different ways, that were used, to analyze the data will be described. For each case the classification results will be shown by using a map of the state of Minnesota, with all the transit systems that were used in the clustering analysis. On section 5.3 the results that were found more meaningful will be presented and will be analyzed more extensively.

In some cases ratios of the raw data, often called as the performance indicators, were used for the clustering analysis. This was done to compare the results by using raw data and by using performance data. On Table 5.1.1 there is a summary of all the performance indicators that were used.

In cases when a pair of variables are high correlated, one of the variables was removed from the set. Seven raw variables and two ratio data were removed because were found to have correlation greater than nine tenths. These variables are presented on Table 5.1.2. Only the first variable in each case was not removed from the data set.

Table 5.1.1 Performance indicators that were used in the clustering analysis.

Miles / Employee	Hours / Employee
Miles / Vehicle	Cost / Mile
Cost / Vehicle	Passengers / Mile
Passengers / Hour	Hour / Population
Passengers / Population	Subsidy / Passenger
Cost / Passenger	Revenue / Cost
Revenue / Subsidy	Miles / Hour

Table 5.1.2 High correlated variables.

Passengers - Vehicles, Employees, Cost, Revenue.
Population - Households, Driver's license.
Sales tax - Property tax.

Cost / Mile - Passenger / Mile, Passenger / Hour.

For the clustering analysis three primary techniques were used:

- One-phase classification.
- Two-phase classification.
- Classification with combination.

5.1.1 One-Phase Classification.

In the one-phase classification all the variables were used in one execution of the algorithm to obtain the results. Different ways to refine the data were used. In some cases the data were scaled, normalized or the high correlated

variables were removed. On the Table 5.1.3 there is a summary of all the one-phase cases that were tested. In all cases the desired number of classes was six (After comparing the classification success index for different number of classes).

Table 5.1.3 Different one-phase classification cases that were tested.

	Cases Tested											
	A:				B:				C:			
	1	2	3	4	1	2	3	4	1	2	3	4
Data Refinement												
Normal	x	x	x	x	.	.	.	.	.	.	.	.
Scaled	.	.	.	.	x	x	x	x	.	.	.	.
Normalized	.	.	.	.	.	.	.	.	x	x	x	x
Variables Used												
All the variables	x	.	x	.	x	.	x	.	x	.	x	.
Exclude the correlated	.	x	.	x	.	x	.	x	.	x	.	x
Type of Variables Used												
Raw data	x	x	.	.	x	x	.	.	x	x	.	.
Ratios	.	.	x	x	.	.	x	x	.	.	x	x

x - applies
. - does not apply

By looking the clustering results for obvious errors the cases A:1 and A:2 were rejected because in both cases there were many misclassifications, e.g. the MTC was assigned with other small transit systems. Then by examining more carefully the results, the cases B:1, B:2, B:3, B:4, C:3 and C:4 were also rejected because there were many outliers and almost all the systems were assigned in one large class. Only the cases A:3, A:4, C:1 and C:2 gave satisfactory results.

In cases A:3 and A:4 the results show three large classes and three small classes. The MTC and Duluth Regular Route were assigned as outliers. These two are much larger than the other systems and are considered as a separate class. Also the system of East Grand Forks was assigned in the second class as outlier. The third small class has three systems Albert Lea, Cottonwood County, Red Wing. Although there was not any obvious error in the results, the three large classes had mixed systems with different characteristics. The results are shown in Fig. 5.1.1.

In cases C:1 and C:2 the results show two large classes, one small class and three outliers. The three outliers are the MTC, the Duluth Regular Route and the Arrowhead. These systems are significantly larger and different than the other systems but also each system is not similar to one another.

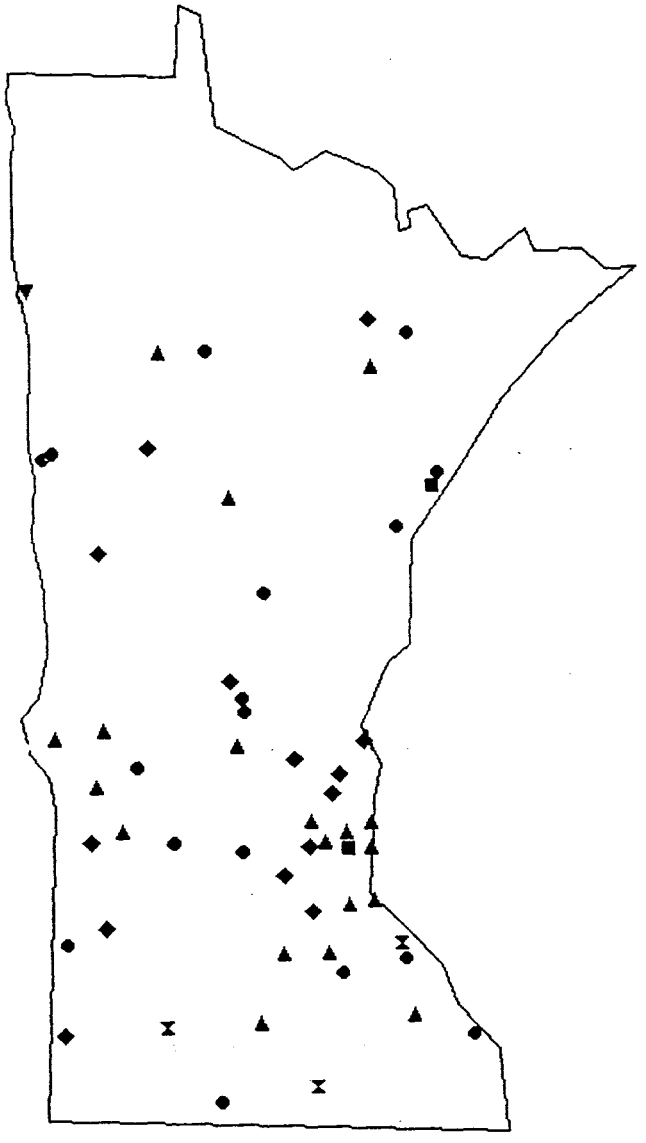

Figure 5.1.1 Classification results for cases A:3 and A:4.

The fourth class consists of four systems which are located close to the Twin Cities Metro Area. They are St. Louis, North Suburban Lines, Medicine Lake Lines and Hopkins. The fifth class consists of systems that are located in the east part of the state of Minnesota and the sixth class consists of systems that are located in the west part of the state (Fig. 5.1.2). The differences between the two classes are basically in the socioeconomic characteristics. Systems in the fifth class are located close to urban areas as the Twin Cities, Duluth, Rochester, St. Cloud. The system of Moorhead, as we can see from Fig. 5.1.2, belongs in the fifth class. This happens because the city of Moorhead is close to the city of Fargo, and for that reason the transit system is more similar with systems that are located close to urban areas.

5.1.2 Two-Phase Classification.

In the two-phase classification the results were obtained in two phases, e.g. classify the systems by using socioeconomic data and then classify each class of systems by using the transportation data only. The data again in some cases were normalized or the high correlated variables were removed. From the previous classification cases it was found that by scaling the data there was not any improvement in the results. On Table 5.1.4 there is a summary of all the two-phase cases that were tested. Always in the first phase the desired

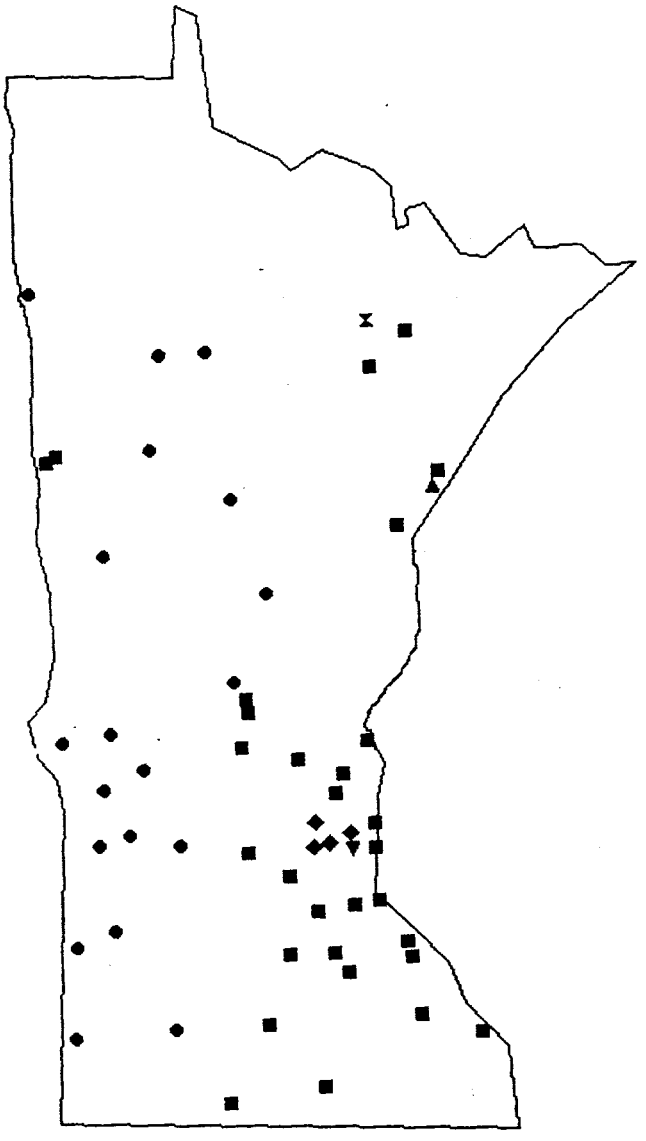

Figure 5.1.2 Classification results for cases C:1 and C:2.

number of classes was six, as in the previous classifications.

Table 5.1.4 Different two-phase classification cases that were tested.

PHASE I	PHASE II
Case S1 Normalized Transportation Raw data	**Case T1** **Case T2**
Case S2 Normalized Transportation Raw data Not correlated	**Case T1** **Case T2**
Case T1 Normalized Socioeconomic Raw data	**Case S1** **Case S2**
Case T2 Normalized Socioeconomic Raw data Not correlated	**Case S1** **Case S2**

By looking the clustering results, we can see that for all the cases where in phase I the transportation data was used, there were some misclassifications and almost all the systems were assigned in the same class. On the other hand, by using the socioeconomic data in phase I, we got satisfactory results. Also there was no significant difference in the results by removing or not the high correlated variables.

In the cases where the socioeconomic data were used in phase I and the transportation data in phase II, the results show two large classes and four small classes. The MTC and the North Suburban Lines were assigned as outliers. These two systems have a significant different service area than the other systems but also the two areas are not similar to one another and are considered as a separate classes. The third class consists of three systems which are located close to the Twin Cities Metro Area. They are St. Louis, Medicine Lake Lines and Hopkins. The fourth class consists of systems that are located close the Duluth Metro Area, St. Louis county. They are Duluth Regular Route, Duluth D.A.R., Arrowhead, Hibbing and Virginia. As in the one-phase classification the fifth class consists of systems that are located in the east part of the state of Minnesota and the sixth class consists of systems that are located in the west part of the state (Fig. 5.1.3). Systems in the fifth class are located close to urban

areas as the Twin Cities, Duluth, Rochester, St. Cloud. Again the system of Moorhead, as we can see from Fig. 5.1.3, belongs in the fifth class. This happens because the city of Moorhead is close to the city of Fargo, and for that reason the transit system is more similar with systems that are located close to urban areas. The classification result of phase II is shown in Fig. 5.1.4. In phase II the two large classes that were defined from the phase I were used to be divided, each one in three small classes, by using transportation data.

5.1.3 Classification with Combination.

From the two previous classification techniques satisfactory results were obtained only when normalized raw data or ratios data were used. A data set was created with the ratios and the normalized data. After the classification method was applied the results show three large and three small classes (Fig. 5.1.5). The first class consists of two systems, the MTC and the Duluth Regular Route. The second class has the system of East Grand Forks and the third class consists of three systems, Albert Lea, Cottonwood county and Red Wing. The other systems are divided in the three large classes. This data set did not give satisfactory results because many systems were misclassified.

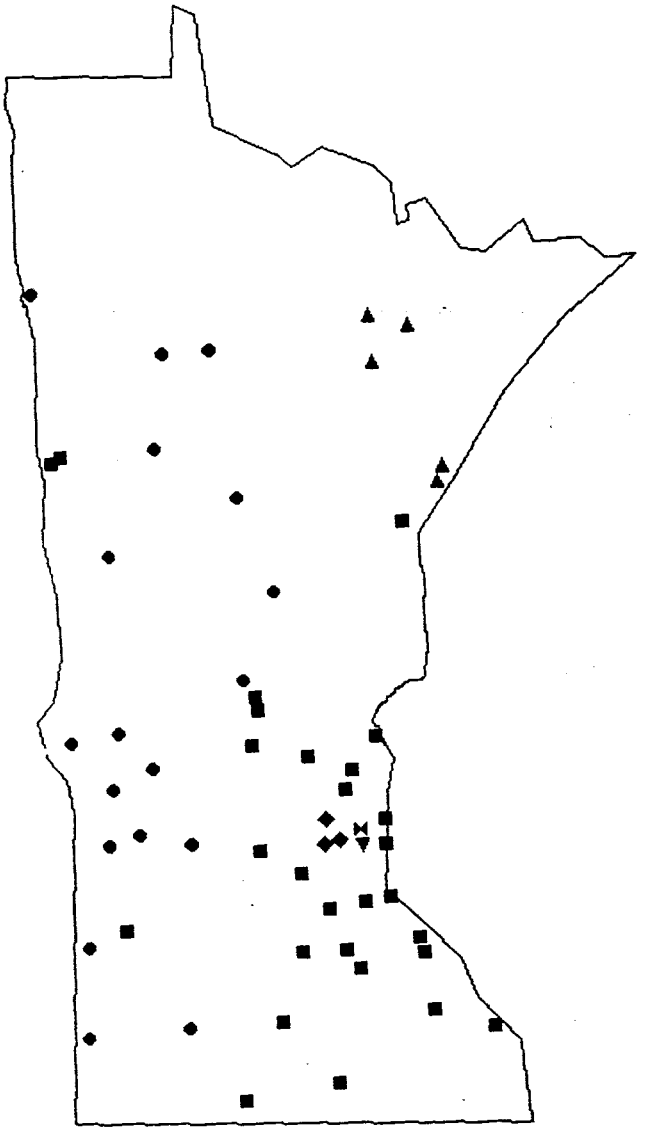

Figure 5.1.3 Results for two-phase classification (Phase I).

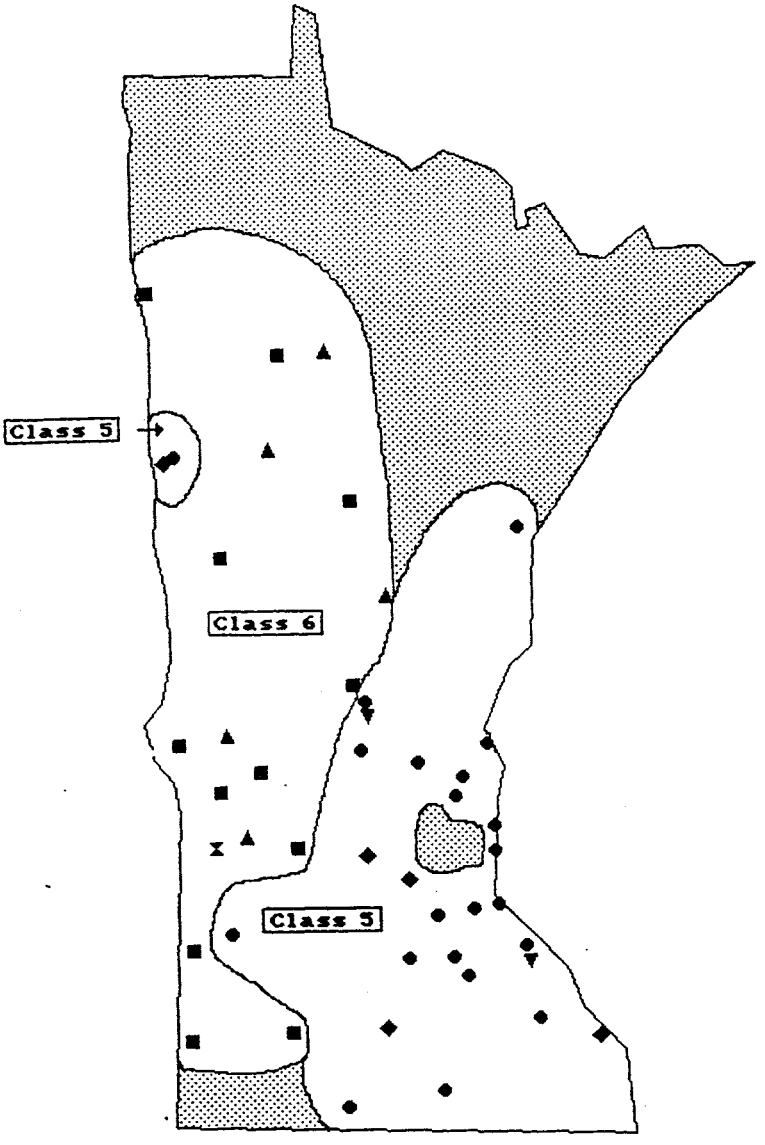

Figure 5.1.4 Results for two-phase classification (Phase II).

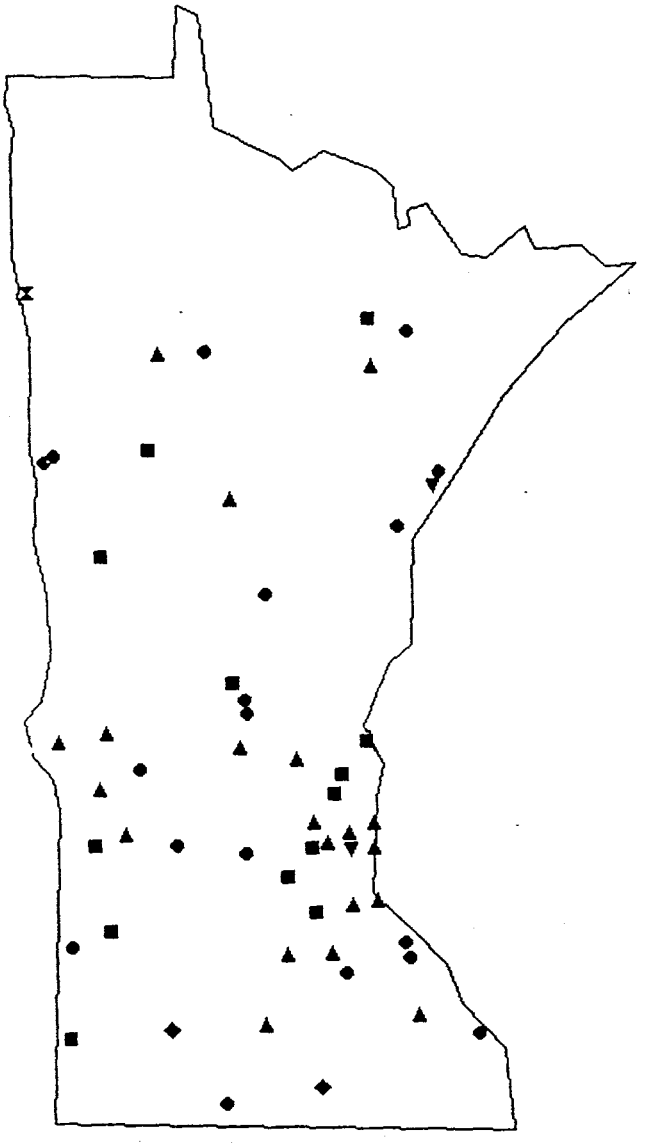

Figure 5.1.5 Results for combination classification.

5.2 Test of the Classification Method.

An ideal experimental design to test the classification method would require a controlled experiment: use the data set of the transit systems, randomly select a part of the data, and then treat the not selected data as a control group for the experiment. The experiment will have two phases.

In the first phase, the training or learning phase, the classifier performs unsupervised classification of the training data set. Then by using the classification results as an input, the feature extractor will reduce the dimensionality of the data set. One way of reducing the number of dimensions is to perform a transformation of variables into a smaller number of orthogonal components. The most efficient way to do this is by extracting eigenvectors through a well known base function, the Karhunen-Loève expansion (principal components analysis), that minimizes a mean-square error criterion. A reduced number of these components will be retained, ensuring that the selected components minimize any loss of information and maintain the differentiation between observed patterns. In the second phase we test the classifier with the remaining data, the verification data.

5.2.1 Classification of the Training Set.

For the first phase of the control experiment a training data set was selected. The classification results will be used as information for the second phase of the experiment. For that reason the training set should be large enough to provide the appropriate information about the data. In this case the 70% of the data was selected as training set, and the 30% of the data as verification set (Table 5.2.1). The two sets were selected randomly. Before the classification method was applied both data sets were normalized.

The new method was applied on the training set for desired number of classes six. Classification in two phases, with socioeconomic and transit indicators, was used because of the good results that this method gave in the previous applications. On Fig. 5.2.1 are the classification results from the first phase (socioeconomic indicators), and on Fig. 5.2.2 are the classification results from the second phase (transit indicators).

By using the classification results, from both phases, the Karhunen-Loève feature extraction was applied to reduce the dimensionality of the data. The eigenvalues and the corresponding eigenvectors are shown in Table 5.2.2 for the socioeconomic indicators and in Table 5.2.3 for the transit

Table 5.2.1 Training set and verification set used in the experiment.

Training set.	
1. Albert Lea	2. Anoka County
3. Appleton	4. Arrowhead
5. Bemidji	6. Benson
7. Brainerd	8. Carver County
9. Chisago County	10. Clearwater County
11. Cloquet	12. Columbia Heights
13. Cottonwood County	14. Dakota County
15. Duluth	16. Duluth D.A.R.
17. East Grand Forks	18. Fairmont
19. Faribault	20. Hastings
21. Hibbing	22. Hopkins
23. Hutchinson	24. Le Sueur
25. Lincoln County	26. Mahube
27. Mankato	28. Marshall
29. Medicine Lake Lines	30. Montevideo
31. Moorhead	32. Moorhead D.A.R.
33. North Suburban Lines	34. Northfield
35. Ortonville	36. Pelican Rapids
37. Red Wing	38. Rochester
39. Twin Cities M.T.C.	

Verification set.	
1. Morris	2. Pine River
3. Pipestone	4. Rochester D.A.R.
5. St. Cloud	6. St. Cloud D.A.R.
7. St. Louis	8. Scott County
9. Sherburne County	10. Tri Cap
11. Tri Valley	12. Upsala
13. Virginia	14. Washington County
15. White Bear Area	16. Willmar
17. Winona	

indicators.

The original variables are listed next to the eigenvectors to indicate the linear combination that makes up each feature.

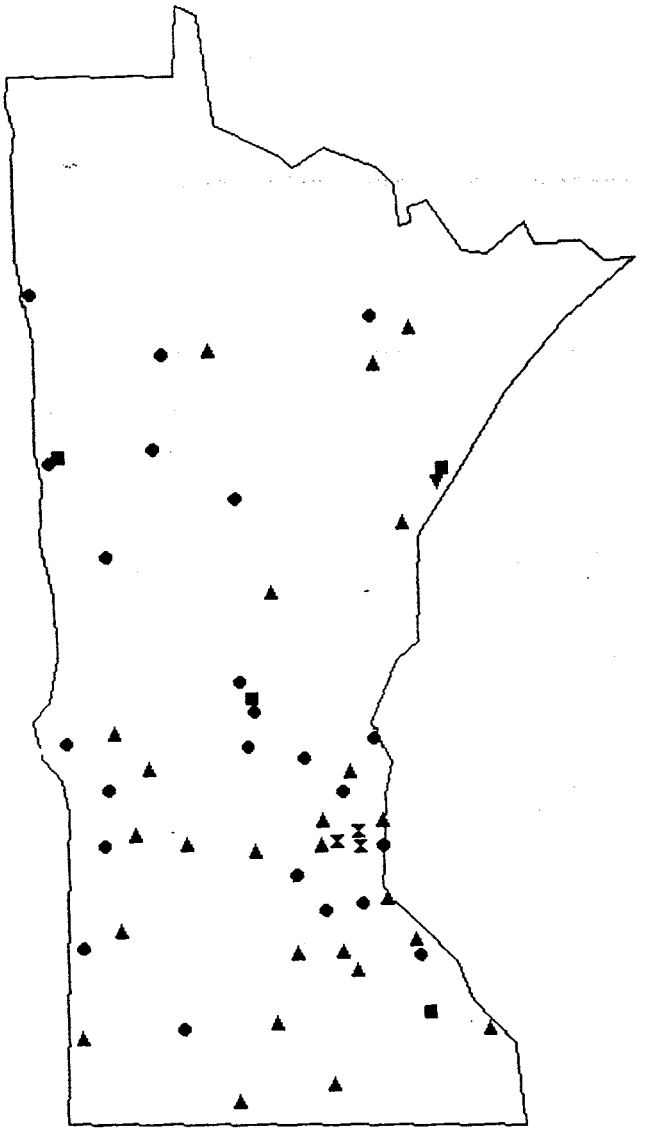

Figure 5.2.1 Classification results from the first phase (socioeconomic indicators).

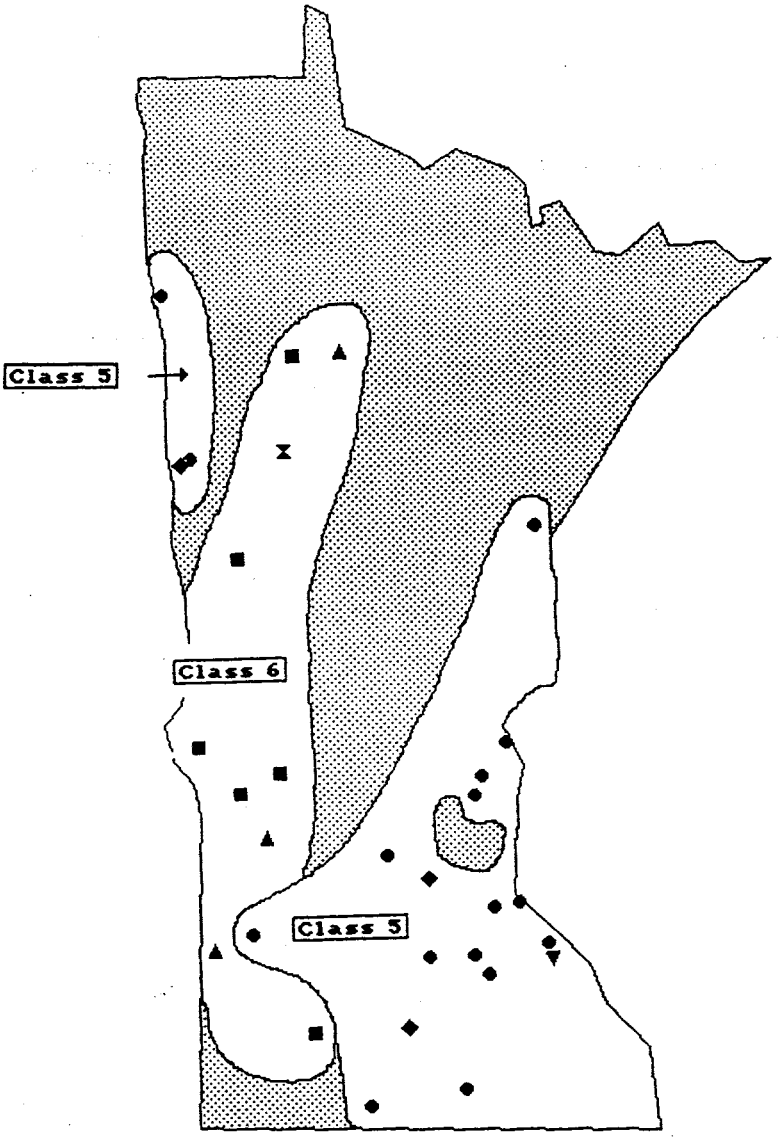

Figure 5.2.2 Classification results from the second phase (transit indicators).

Table 5.2.2 Eigenvectors for socioeconomic data.

	(1)	(2)	(3)	(4)
Eigenvalues	19.7186	1.9929	1.7027	1.0663
% of Variation	76.89	7.77	6.63	4.15
Cumulative %	76.89	84.66	91.29	95.44
Population	-0.5795	-0.4198	-0.6630	0.0991
Income	-0.2105	0.2146	-0.0350	0.2678
Property Tax	-0.4250	-0.0280	0.4525	-0.4790
Hwy Expenditures	-0.4026	-0.4123	0.4236	-0.0501
Population Density	-0.5084	0.7263	-0.0198	0.0712
Age	0.0071	-0.1224	0.0721	-0.1992
Unemployment	-0.0192	-0.2409	0.3296	0.6675
Poverty Level	0.1342	-0.0783	-2.4600	-0.4427

	(5)	(6)	(7)	(8)
Eigenvalues	0.7154	0.2424	0.1362	0.0698
% of Variation	2.78	0.94	0.53	0.27
Cumulative %	98.22	99.16	99.69	99.96
Population	0.0299	-0.0181	-0.1876	0.0436
Income	-0.5449	-0.3015	0.2317	-0.6284
Property Tax	-0.1995	0.0835	-0.5558	-0.1685
Hwy Expenditures	0.0070	0.0640	0.6751	0.1610
Population Density	0.4125	0.1317	0.1236	0.0745
Age	0.5509	-0.7578	-0.0049	-0.2492
Unemployment	0.3563	0.2687	-0.2630	-0.3447
Poverty Level	0.2478	0.4834	0.2487	-0.6018

Table 5.2.3 Eigenvectors for transportation data.

Class 5.

	(1)	(2)	(3)	(4)
Eigenvalues	0.2572	0.0059	5.4E-4	5.9E-6
% of Variation	97.55	2.23	0.20	0.00
Cumulative %	97.55	99.78	99.98	99.99
Government Subsidy	-0.3756	-0.6575	-0.6059	0.2435
Vehicle Hours	-0.6422	0.1016	-0.0174	-0.7595
Vehicle Miles	-0.6057	0.5399	0.0242	0.5838
Passengers	-0.2819	-0.5154	0.7949	0.1511

Class 6.

	(1)	(2)	(3)	(4)
Eigenvalues	0.6644	0.0410	3.6E-3	3.5E-5
% of Variation	93.70	5.78	0.50	0.00
Cumulative %	93.70	99.48	99.98	99.99
Government Subsidy	0.0380	-0.7396	-0.5940	0.3139
Vehicle Hours	-0.5848	-0.3166	-0.0374	-0.7458
Vehicle Miles	-0.8099	0.1801	0.0185	0.5577
Passengers	0.0196	-0.5658	0.8033	0.1845

As an example, in the Table 5.2.3 the first feature, for class 5, would be computed by the equation:

$$\begin{aligned} Y1 = & -0.3756\,(\textit{Government subsidy}) \\ & -0.6422\,(\textit{Vehicle Hours}) \\ & -0.6057\,(\textit{Vehicle Miles}) \\ & -2819\,(\textit{Passengers}) \end{aligned} \qquad (4.4.1)$$

This feature alone describes 97.55 percent of the variation in the original data.

From the features on Table 5.2.2 the first five were used in the second face of the experiment. According to the eigenvalues the first five features describe 98.22% of the variation. From the features on Table 5.2.3 only the first was used in the experiment for both classes. According to the eigenvalues the first feature in class 5 describes the 97.55% of the variation and the first feature in class 6 describes the 93.70% of the variation.

5.2.2 Test the Classifier by Using the Verification Set.

In the second phase of the control experiment each transit system in the verification set was classified by using supervised classification. The information that was obtained from the first phase of the experiment was used in this phase. The results are presented on Fig. 5.2.3. By comparing the results from this classification with the results from the classification on section 5.1 we can see that only two transit systems were misclassified.

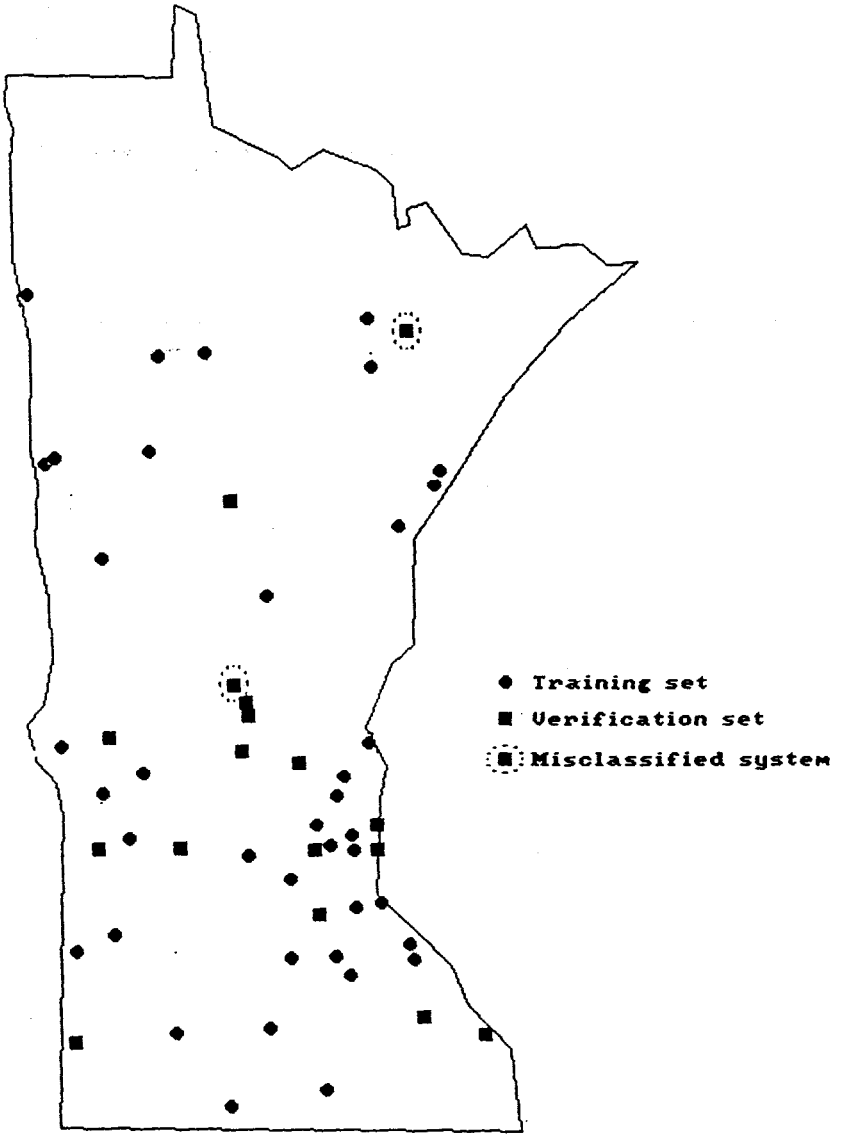

Figure 5.2.3 Classification results of the verification data set.

5.3 Recommended Classification.

The cases that gave the best results will be analyzed in this chapter. There is no specific set of rules that can be directly applied to determine if a classification result is good or not. Three characteristics of the resulting classifications will be considered:

- How the clusters look when graphed.
- The variation within each group as it compares to the variation between groups.
- How the classes compare to present knowledge of the systems and peer groupings presently in use.

Scatter plots of one variable versus a second variable show the geometric structure of the class. By looking at these plots one can see if the classes form valid clusters in the data space. Scatter plots of the first and second features from the Karhunen-Loève transformation will be considered for the two data sets. The first two features describe usually over ninety percent of the variation, so they should give a good indication as to what the classes look like.

A comparison of the variation within groups to the variation between groups will not only indicate how similar the individual members of a group are, but also will tell if the

groups are distinct. Two groups may have a small within-group variance indicating that the members within each group are similar to one another. However, the two groups may actually be one cluster that has been incorrectly divided. A high variation between groups will confirm that these two groups are correctly separated.

In addition to the scatter plots and the variation within and between each group, knowledge of the systems being clustered will be used. The classes will be compared to present peer groupings to see if there is any resemblance. Also, knowledge about each individual system will help determine whether or not it is appropriate to put two particular systems in the same class.

5.4 Scatter Plots.

There are two classifications that gave satisfactory results:
a) When the ratios were used as a data set and the classification was made in one phase,
b) When the classification was made in two phases, with socioeconomic and transportation data.

In each case the Karhunen-Loève feature extraction was used to

define the most significant features. For each scatter plot the features that correspond to the two largest eigenvalues were used.

5.4.1 One-Phase Classification (Ratios).

Twelve ratio data were used in this case and the classification was done in one-phase. On Fig. 5.4.1 there is a scatter plot of the first versus the second most significant feature. These two features represent the 88.11% of the variation that we can have from all the data.

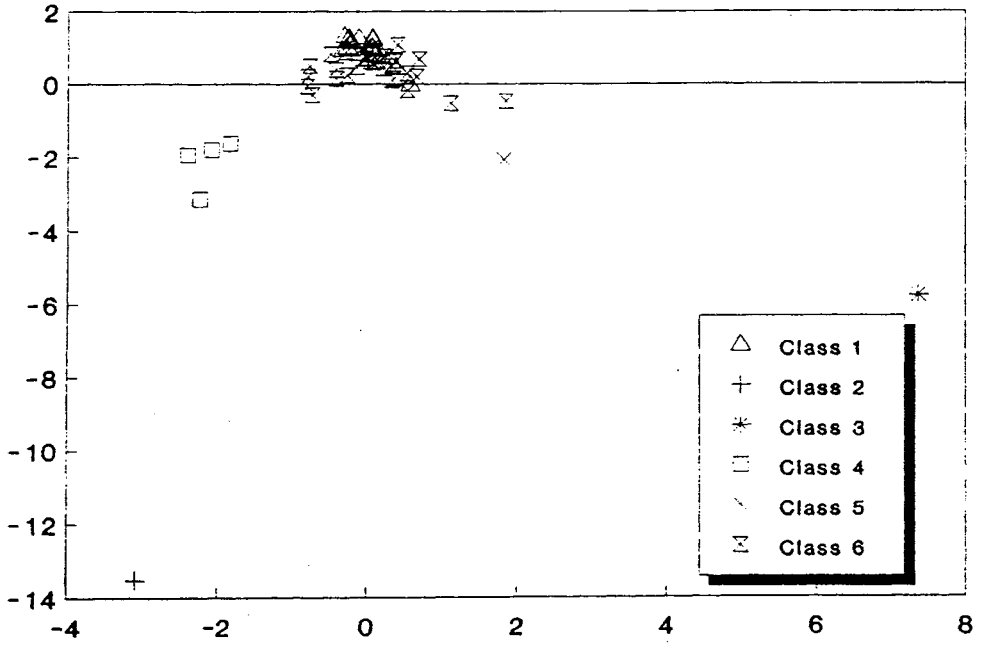

Figure 5.4.1 Scatter plot for one-phase classification.

Four of the six classes can be separated from one another and from the other two classes, there is no overlap among those classes. Two classes, the first and the sixth, have overlap on these two features. To understand the way that these two classes were divided we need to have a graph from the other features also.

Three classes have only one member in each, Twin Cities M.T.C., Duluth Regular Route, and Arrowhead. From the scatter plot it is obvious that Twin Cities M.T.C. (Class 2), and Duluth Regular Route (Class 3), are very different from the other systems. Arrowhead is the only member of the fifth class even though we can not see any significant difference on the plot from the systems in class 6. The reason is that by looking this graph, we do not take into account the other features. The influence of the other features, makes the system of Arrowhead to be assigned in a different class. Class 4 includes four systems that are located in the Twin Cities Metro Area the systems are: Hopkins, Medicine Lake Lines, St. Louis Park, and North Suburban Lines.

5.4.2 Two-Phase Classification (Socioeconomic, Transportation).

In this case eight socioeconomic data were used in the first phase, and four transportation data were used in the second

phase. On Fig. 5.4.2 there is a scatter plot of the first versus the second most significant feature from the socioeconomic data. These two features represent the 84.64% of the variation that we can have by using all the socioeconomic data. The scatter plot for this classification

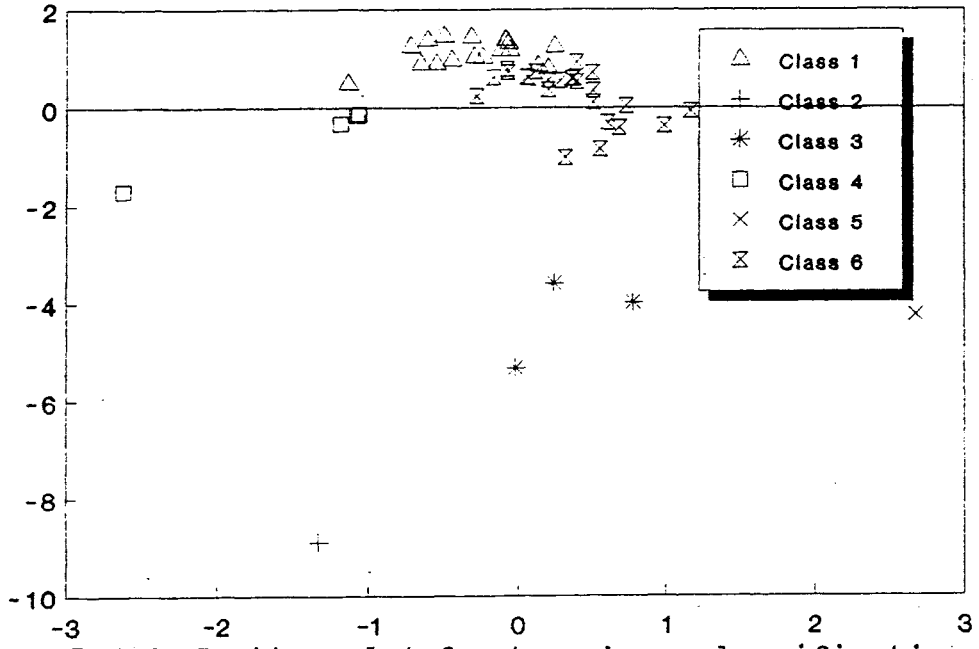

Figure 5.4.2 Scatter plot for two-phase classification (Phase I).

shows that the classes are well separated. There is no overlap among the classes but two of them are very close together.

Two classes have only one member in each, Twin Cities M.T.C., and Medicine Lake Lines. From the scatter plot it is obvious that Twin Cities M.T.C. (Class 2), and North Suburban Lines (Class 5), are very different from the other systems. Class

3 includes systems that are located in the Twin Cities Metro Area: Hopkins, Medicine Lake Lines, and St. Louis Park. Class 4 consists with systems that are located close to the City of Duluth Metro Area: Duluth Regular Route, Duluth D.A.R., Arrowhead, Hibbing, and Virginia.

In the second phase of the classification, four transportation characteristics were used to classify the systems in classes 1 and 6. On Fig. 5.4.3 there is a scatter plot of the first versus the second most significant feature for class 1. These two features represent the 99.76% of the variation that we can have by using all the transportation data. From the scatter plot we can see that there are two large classes and one class that has one member.

On Fig. 5.4.4 there is a scatter plot of the first versus the second most significant feature for class 6. The two features represent the 99.75% of the variation that we can have by using all the transportation data. From the scatter plot we can see that there are three classes well separated. Class 2 includes two systems that are quite different from the other systems: Rochester and St. Cloud

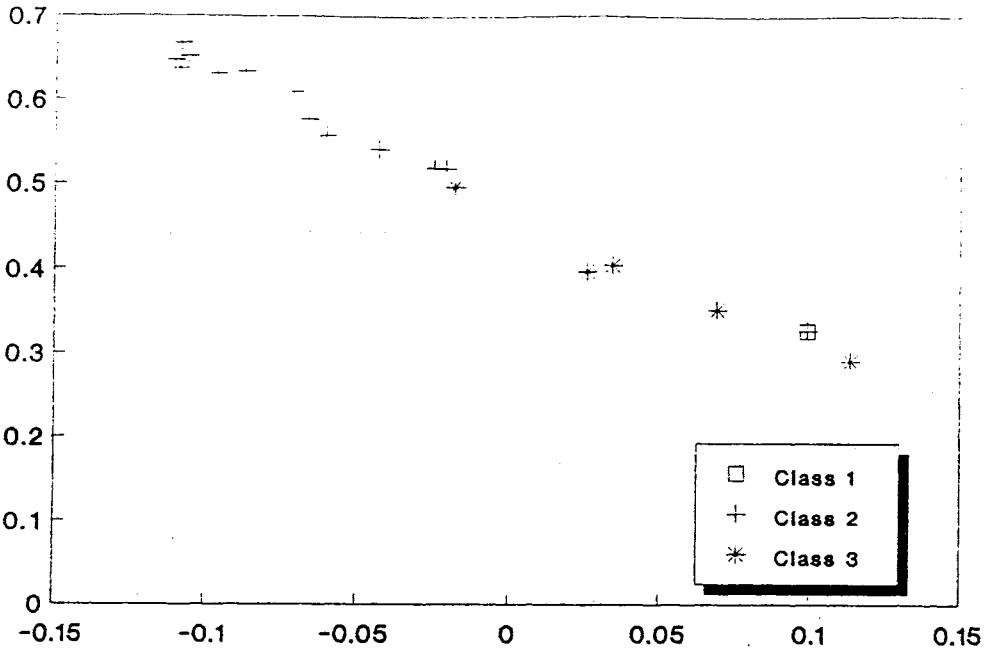

Figure 5.4.3 Scatter plot for two-phase classification (Phase II, Class 1).

5.5 Existing Classifications (MN/DOT).

All transit providers in the state of Minnesota that apply for assistance from the state are included in the analysis. The Minnesota Department of Transportation and, in more recent years, the Regional Transit Board, are responsible for administering state funds and assisting in planning for these systems. MN/DOT currently uses a peer grouping based on population and type of service provided. This classification is presented on Fig. 5.5.1.

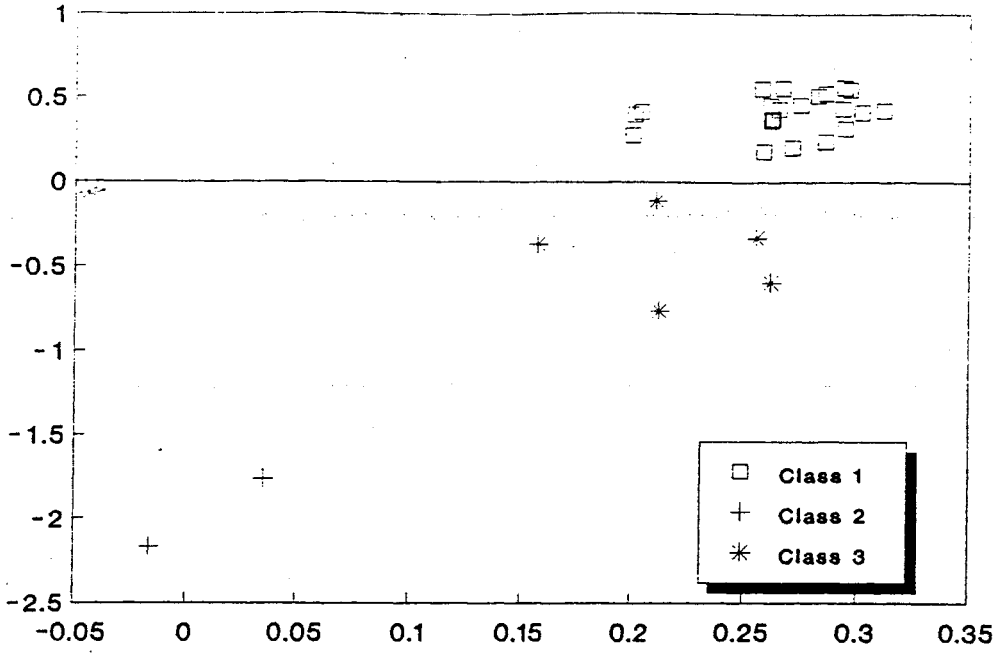

Figure 5.4.4 Scatter plot for two-phase classification (Phase II, Class 6).

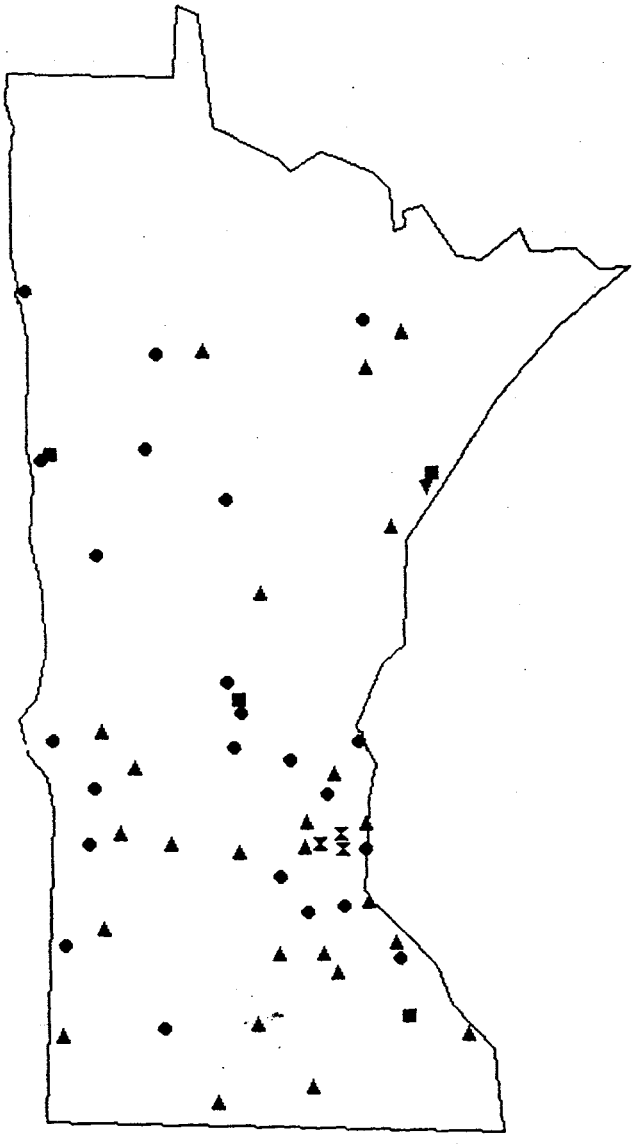

Figure 5.5.1 Classification of the transit systems based on population and type of service (MN / DOT).

CONCLUSIONS

6 Conclusions.

A set of peer groups of Minnesota transit systems has been found by using a clustering analysis method. The results of the different classification approaches were compared to determine the most useful classification. For all the cases the new proposed method was used.

The classification method that is proposed by this study gives significantly better results than any other method found in the literature because of the advantages that has over the other methods. The new method does not require definition of any parameters or of any initial settings and always gives a unique result for a given data set. The classification result is not depended on the geometric properties of the data and the method allows reassignment of objects in order to obtain the best classification. The classification results can be used as a standard point of reference, e.g. between transit manager and MN/DOT, because they are unique.

The recommended final classification was chosen on the basis of its performance on three criteria. First of all, the scatter plots show distinct clusters that are reasonably uniform in size and shape. Also, the classification success

index gets the maximum value for this classification approach. A high value of this index indicates small variation within each group relative to the variation between groups. Finally, the classes show some similarity to the peer groups formed by MN/DOT. Some of the peer groups were merged while others were divided.

The classification in two phases, with socioeconomic and transportation data, led to peer groups whose members are similar in service area size and type of service provided. This was the basis for the original MN/DOT peer groups, although they did not adhere fully to their population criteria.

A larger set of characteristics was used in this analysis compared to previous research. This was possible in part because of the use of Karhunen-Loève expansion. Since more characteristics were used, the members of each peer group now have more characteristics in common. Therefore, they are more likely to have similar goals and objectives. Consequently, it is more reasonable to compare the members within a group. By nature of the clustering algorithm, there are no outliers in any of the groups. Instead outliers are assigned to separate groups.

The recommended peer groups can be used, by MN/DOT and other transit agencies, for performance evaluation. Funding allocation and managerial assistance can be determined on the basis of these groups. Similarly, scheduling and fare policies can be used for systems in the same peer groups.

The classification method does not give direct answers to some common problems the transit manager faces but the method can be a useful part of an expert system for optimal design of transit service. This expert system will use the classification results and data from the transit manager, the MN/DOT, the Census and the transit data to decide on designing of a new system or changing the existing transit services. Also it may give answers in other requirements from managers and MN/DOT, e.g. in many cases two systems with the same socioeconomic and operating characteristics have different performance without any obvious reason.

For further analysis, other characteristics should be considered. More data may be obtained from MN/DOT, the transit operators, and other sources, as to what characteristics they feel are important in determining peer groups. Changes in these characteristics over time could be used in the clustering analysis. The use of the Karhunen-Loève expansion makes this type of analysis more reasonable by

reducing the dimensionality of the data without any significant loss of the original information.

BIBLIOGRAPHY

Bibliography.

1. Anderberg M. R. "Cluster analysis for applications", New York, Academic Press, 1973.

2. Anderson S. C., and Fielding G. J. (1982). "Comparative analysis of transit performance", University of California, Irvine. Institute of Transportation Studies. Final Report, UMTA-CA-11-0020-1.

3. Baker F. B., and Hubert L. J. (1975). "Measuring the power of hierarchical cluster analysis", Journal of the American Statistical Association, 70, 31-38.

4. Ball G. H. (1965). "Data analysis in the social sciences: What about the details? In W. R. Rector (Chairman)", American Federation of Information Processing Societies, (Fall Joint Computer Conference), 27, 533-539.

5. Bartko J. J., Strauss J. S., and Carpenter W. T. (1971). "An evaluation of taxometric techniques for psychiatric data", Classification Society Bulletin, 2, 2-28.

6. Bezdek J. C. (1974). "Numerical taxonomy with fuzzy sets", Journal of Mathematical Biology, 1, 57-71. (a)

7. Bezdek J. C. (1974). "Cluster validity with fuzzy sets", Journal of Cybernetics, 3, 58-73. (b)

8. Blashfield R. K. (1976). "Mixture model tests of cluster analysis: Accuracy of four hierarchical agglomerative methods", Psychological Bulletin, 83, 377-388. (a)

9. Blashfield R. K. (1976). "Questionnaire on cluster analysis software", Classification Society Bulletin, 3, 25-42. (b)

10. Blasfield R. K., and Aldenderfer M. S. (1978). "The literature on cluster analysis, Multivariate Behavioral Research, 13, 271-295.

11. Bottiny W. H., and Goley B. T. (1967). "A classification of urbanized areas of transportation analysis", Highway Research Record, No. 194, pp. 32-61. Transportation Research Board, Washington, DC.

12. Calinski R. B., and Harabasz J. (1974). "A dendrite method for cluster analysis", Communications in Statistics, 3, 1-27.

13. Carlson K. A. (1972). "Classes of adult offenders: A multivariate approach", Journal of Abnormal Psychology, 80, 84-93.

14. Clifford H. T., and Stephenson W. "An introduction to numerical classification", New York, Academic Press, 1975.

15. Cunningham K. M., and Ogilive J. C. (1972). "Evaluation of hierarchical grouping techniques-a preliminary study", Computer Journal, 15, 209-215.

16. D' Andrade, R. G. (1978). "U - statistic hierarchical clustering", Psychometrika, 43, 59-67.

17. Dubes R., and Jain A. K. (1979). "Validity studies in clustering methodologies", Pattern Recognition, 11, 235-254.

18. Edwards A. W. F. and Cavalli-Sforza L. L. (1965). "A method of cluster analysis", Biometrics, 21, 362-375.

19. Everitt B. S. (1979). "Unresolved problems in cluster analysis", Biometrics, 35, 169-181.

20. Fielding G. J., Brenner M. E., and Faust K. "Typology for bus transit", Transportation Research, Vol. 19A, 1985.

21. Fisher L., and Van Ness J. W. (1971). "Admissible clustering procedures", Biometrika, 58, 91-104.

22. Friedman H. P., and Rubin J. (1967). "On some invariant criteria for grouping data", Journal of the American Statistical Association, 62, 1159-1178.

23. Golob T. F., Canty E. T., and Gustafson R. L. (1972). "Classification of metropolitan areas for the study of new systems of arterial transportation", General Motors Research Laboratory, Transportation Research Department, Michigan, Res. publ. GMR-1225.

24. Jardine N., and Sibson R. "Mathematical taxonomy", New York: Wiley, 1971.

25. Jones K. J. (1968). "Problems of grouping individuals and the method of modality", Behavioral Science, 13, 496-511.

26. Kneafsey J. T. (1975). "Transport Economic Analysis", Lexington Books, Lexington, MA.

27. Lance G. N., and Williams W. T. (1967). "A general theory of classifactory sorting strategies. I. Hierarchical systems", Computer Journal, 9, 373-380. (a)

28. Lance G. N., and Williams W. T. (1967). "A general theory of classifactory sorting strategies. II. Clustering systems", Computer Journal, 10, 271-277. (b)

29. Lorr M., and Radhakrishnan B. K. (1967). "A comparison of two methods of cluster analysis", Educational and Psychological Measurement, 27, 47-53.

30. McQuitty L. L. (1967). "A mutual development of some typological theories and pattern-analytic methods", Educational and Psychological Measurement, 27, 41-46.

31. Miller J. H. (1980). "The use of performance-based methodologies for the allocation of transit operating funds", Transportation Quarterly, 34, 574.

32. Milligan G. W. (1980). "An examination of the effect of six types of error perturbation on fifteen clustering algorithms", Psychometrica, 45, 325-342.

33. Milligan G. W. (1981). "A review of Monte Carlo tests of cluster analysis", Multivariate Behavioral Research, 16, 379-407. (a)

34. Milligan G. W. (1981). "A discussion of procedures for determining the number of clusters in a data set", Paper presented at the meeting of the Classification Society, Toronto. (b)

35. Nash C. A. (1980). "Management objectives, fares and service levels in bus transport", Journal of Transportation Economic Policy, 12, 70-85.

36. Nelson G. R. (1972). "An economic model of urban bus transit operations", Unpublished Ph.D. dissertation, Rice Univ., Houston, TX.

37. Patton T. A. (1983). "Transit performance indicators", U. S. Department of Transportation, Transportation Systems Center, Cambridge. Staff Study (ss-67-0.3-01).

38. Peay E. R. (1975). "Nonmetric grouping: Clusters and Cliques", Psychometrika, 40, 297-313

39. Peebles C. (1972). "Monothetic-divisive analysis of

Moundville burials", Newsletter of Computer Archaeology, 1972.

40. Purcher J. A., Markstedt A., and Hirschman I. (1983). "Impacts of subsidies on the costs of urban public transport", Journal of Transport Economic Policy, 17, 155-176.

41. Sibson R. (1971). "Some observations on a paper by Lance and Williams", Computer Journal, 14, 156-157.

42. Sinha K. C., Jukins D. P., and Bevilacqua O. M. (1980). "Stratification approach to evaluation of urban transit performance", Transportation Research Record, 761, 20-27.

43. Sinha K. C., and Kumaras C. "Development of an approach for the allocation of the public mass transit fund of Indiana", School of Civil Engineering, Purdue University, West Lafayette, IN, July 1985.

44. Sneath P. H. A. (1966). "A comparison of different clustering methods as applied to randomly-spaced points", Classification Society Bulletin, 1, 2-18.

45. Sneath P. H. A., and Sokal R. R. "Numerical taxonomy", San Francisco, W. H. Freeman, 1973.

46. Sokal R. R., and Rohlf F. J. (1962). "The comparison of dendrograms by objective methods", Taxon, 11, 33-40.

47. Tardif T. J., Mohammadi B., and Vaziri M. (1977). "Analysis of the Sacramento area transportation market", California Department of Transportation, Sacramento, CA.

48. Tou J. T., and Gonzales R. C. "Pattern recognition principles", Addison - Wesley, Reading, MA, 1974.

49. Tryon R. C. "Cluster analysis", Ann Arbor: Edward Brothers, 1939.

50. Tryon R. C., and Bailey D. E. "Cluster analysis", New York: McGraw-Hill, 1970.

51. Veatch J. F. (1973). "Cost and demand for urban bus transit", Unpublished Ph.D. dissertation, Univ. Illinois at Urbana-Champaign.

52. Von Eye A., and Wirsing M. (1978). "An attempt for a mathematical foundation and evaluation of MACS, a method

for multidimensional automatical cluster detection", Biometrical Journal, 20, 655-666.

53. Von Eye A., and Wirsing M. (1980). "Cluster search by enveloping space density maxima", In M. M. Barritt and D. Wishart (Eds.), COMPSTAT 1980. Vienna: Physica-Verlang.

54. Ward J. H. (1963). "Hierarchical grouping to optimize an objective function", Journal of American Statistical Association, 58, 236-244.

55. Watanabe Satoshi "Pattern recognition - human and mechanical", John Wiley and Sons, 1985.

56. Whallon R. (1972). "Anew approach to pottery typology", American Antiquity, 37, 13-34.

57. Wishart D. (1969). "Mode analysis: A generalization of nearest neighbor which reduces chaining effects. In A. J. Cole (ed.)", Numerical Taxonomy, London: Academic Press.

58. Young T. Y., and Galvert T. W. "Classification, estimation and pattern recognition", American Elsevier Publishing Company Inc., New York, 1974.

APPENDIX A

A Classification Success Index.

A method for identifying the number of clusters of points in a multidimensional Euclidean space is described and an informal indicator of the "best number" of clusters is suggested. It is a "variance ration criterion" giving some insight into the structure of the points. A familiar objective function applicable in cluster analysis is the within-group (cluster) sum of squares (WGSS). It seems natural to regard the optimal grouping of n points into k clusters as that for which WGSS is minimized. This criterion reflects a desire to find some minimum variance spherical clusters.

A.1 Methodology Used.

Suppose there are n individuals (or samples from n populations) with observations on the same v variables for each individual. We may imagine them as being represented by n points in a v-dimensional Euclidean space, P_1, P_2, ..., P_n. The variables permit the computation of an n by n distance matrix. If we denote the original v by n data matrix by $\underline{X}$, with rows given by the observed variables and with columns

given by the individuals, we can write $\underline{X} = (\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n)$, where the column $\underline{x}_i$ is a vector of the v coordinates of the point P_i. If we refer the coordinates to orthogonal axes of an ordinary Euclidean space then the distance d_{ij} between P_i and P_j will be properly defined by the function:

$$d_{ij}^2 = (x_i - x_j)'(x_i - x_j), \quad i,j = 1,2,\ldots,n \qquad (A.1.1)$$

A similar formula applies to the distance between a point and the centroid of the n points. In the approach to cluster analysis which we follow, the dispersion of a group of n points is measured by the sum of the squared distances of the points from their centroid. This sum is equal to the trace of the matrix $\underline{R}$, but may be obtained from the pairwise distances dij by applying the formula:

$$Trace\ \underline{R} = n^{-1}(d_{12}^2 + d_{13}^2 + \ldots + d_{n-1,n}^2) \qquad (A.1.2)$$

If we examine a split leading to a division of the n points into k groups of $n_1, n_2, \ldots, n_k$ points ($n_1 + n_2 + \ldots + n_k = n$), then the WGSS is calculated by applying the right hand side of Eq. A.1.2 to each of the clusters separately and then summing the results. We may then write:

$$WGSS = Trace\ \underline{W} = Trace\ R_1 + Trace\ R_2 + \ldots + Trace\ R_k \qquad (A.1.3)$$

where

$$Trace\ R_g = n_g^{-1}(d_{12}^2(g) + d_{13}^2 g + \ldots + d_{n_g-1,n_g}^2(g)) \quad (A.1.4)$$

with $d_{ij}(g)$ denoting the distance between points P_i and P_j in the gth cluster (g = 1, 2, ..., k). If k, the number of clusters, is not known, we proceed as follows: first we take k = 2, then k = 3, and so on. At each stage we find "the best sum of squares split", for which we calculate not only the (minimum) WGSS, but also the (maximum) BGSS and the variance ratio criterion:

$$VRC = \frac{\frac{BGSS}{k-1}}{\frac{WGSS}{n-k}} \quad (A.1.5)$$

Eq. A.1.5 give an informal indicator for the best number of groups. It is evident that this criterion is analogous to the F-statistic in univariate analysis.

When between two classification results the number of classes is the same, then one way of comparing the two results is to compare the variation within the groups to the variation between groups, by considering the matrix BW^{-1}. This matrix is the result of multiplying B, the matrix of the sum of squares between groups by the inverse of W, the matrix of the sum of squares within groups. The size of BW^{-1} indicates the general relation between the variation within groups and the

variation between groups. A large matrix is desirable as it suggests a high variation between groups compared to variation within groups. In this case, the determinant of BW^{-1} is used as an indicator of the size of this matrix. This index can not be used in cases where the two classification results do not have the same number of classes, because the determinant tends to increase as the number of classes increases.

In order to test the sensitivity of the index a small data set was used. The data set consists of nine points and there are three classes. The first class has four points, the second class has three points and the third class has two points (Fig. A.1.1). The index was tested for two different cases.

In the first case the position of one class was changing. In this way the variation within the groups was the same but the variation between the groups was changing. In that test the class A was moving towards to the class B and after a position change the index was calculated. In Fig. A.1.2 there is a line graph which shows all the values of the index. Line graph "a" describes the variation between classes A and B, and line graph "b" describes the variation between classes A and C. On the figure we can see that the variation between A and B gets the minimum value when the distance between the two classes becomes almost zero. Similarly, the variation between

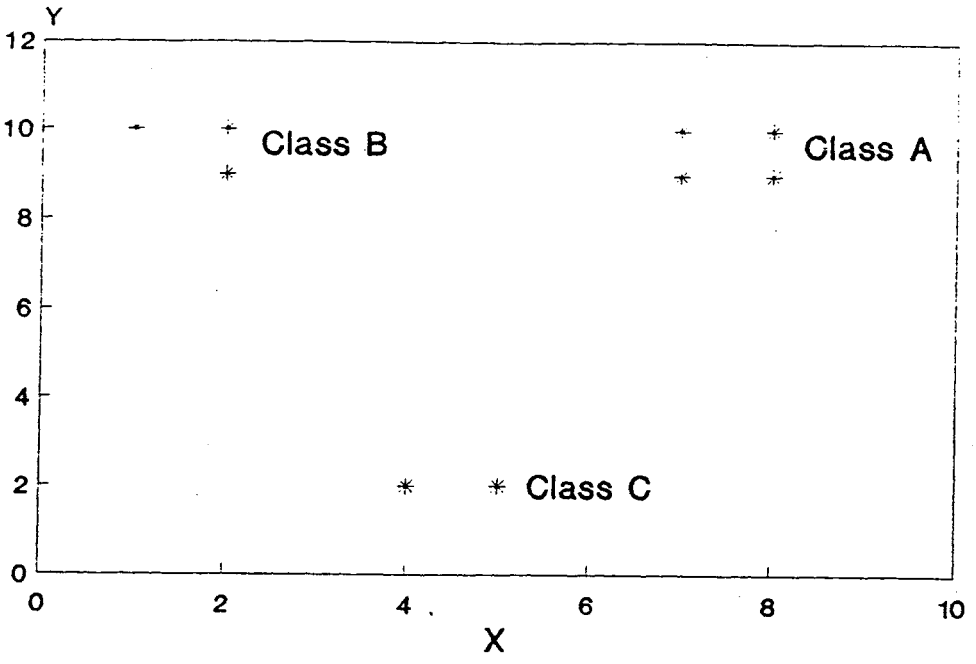

Figure A.1.1 Data set that was used for testing of the classification index.

classes A and C gets the minimum value when the distance between the two classes has the minimum value.

In the second case the positions of the cluster centers remained the same but the two points in class C become more spread out. In this way the variation between the groups was the same but the variation within the groups was changing. In Fig. A.1.3 there is a line graph which shows how the index was changing. In the beginning the index is very sensitive and for the first steps the value is reduced to the 10% of the initial value, then the differences are small.

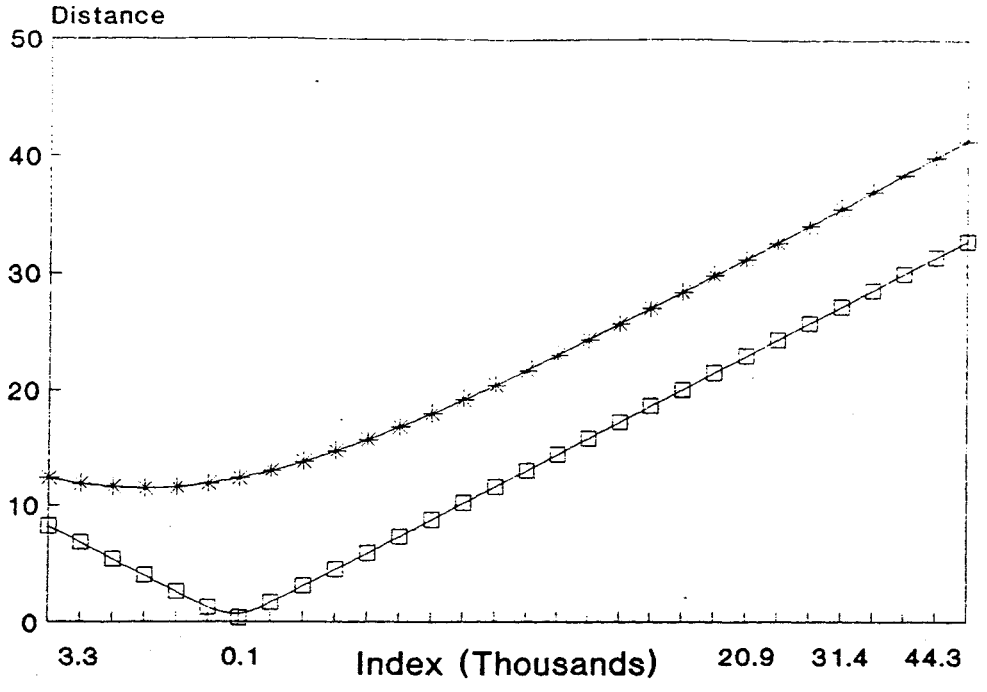

Figure A.1.2 Change of the classification index as the class A is moving.

The two testing cases show that the classification index gives good results and that is sensitive in changes of the within group variation and in changes of the between group variation. Because of the differences between the two classification indexes each index was used for a different purpose. The first classification index was used to specify the optimum number of classes and the second classification index was used in order to decide for the most appropriate classification method.

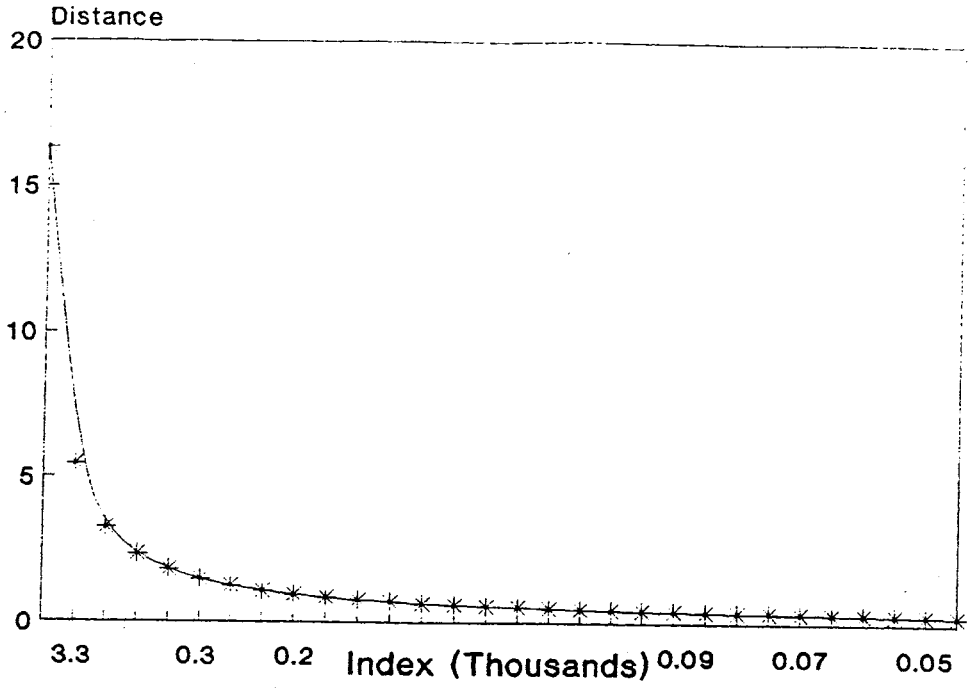

Figure A.1.3 Change of the classification index as the class C is spread out.

APPENDIX B

B Classification and Testing Programs.

In this research several classification methods were tested and many data sets were used, real and randomly generated. Also different ways to refine the data were used, the data were scaled or normalized. In order to do this work fast and more accurately two computer programs were written for these purposes. The first program was written to test the different classification algorithms by using real or randomly generated data. The second program is more complicated because it can handle data bases, apply classification methods, analyze and represent the classification results. In this section the structure of these two programs will be illustrated, and a summary of their functions will be discussed. Both programs were written in Microsoft Quick C, v5.1 and a 386 PC Compatible machine was used. Also a EGA or VGA color monitor is required to have graphic representations of the data and the classification results.

B.1 Classification Program.

This program was used to apply the classification algorithms on real data. Although data from transit system were used in

this research, the program was written to have general use. The functions that are included in this program can be divided in three categories:

- ♦ Data base functions.
- ♦ Display data functions.
- ♦ Classification functions.

The structure of the program is illustrated on Fig. B.1.1.

B.1.1 Data Base Functions.

These functions give the ability to the user to deal with large data sets. In this study the largest data set had 56 transit systems and for each system 33 indicators were included. The data base functions are the following:

- ♦ New. This function creates a new data base and prompts the user to input the data. Then a file is being created with all the data.
- ♦ Open. This function restores a data base from a file.
- ♦ Edit. This function gives the ability to the user to change any value in the data base.
- ♦ Add, Delete. With these two functions the user can add/delete a system in the data base or to add/delete an indicator in/from all the systems.
- ♦ Transform. The user can create new indicators by

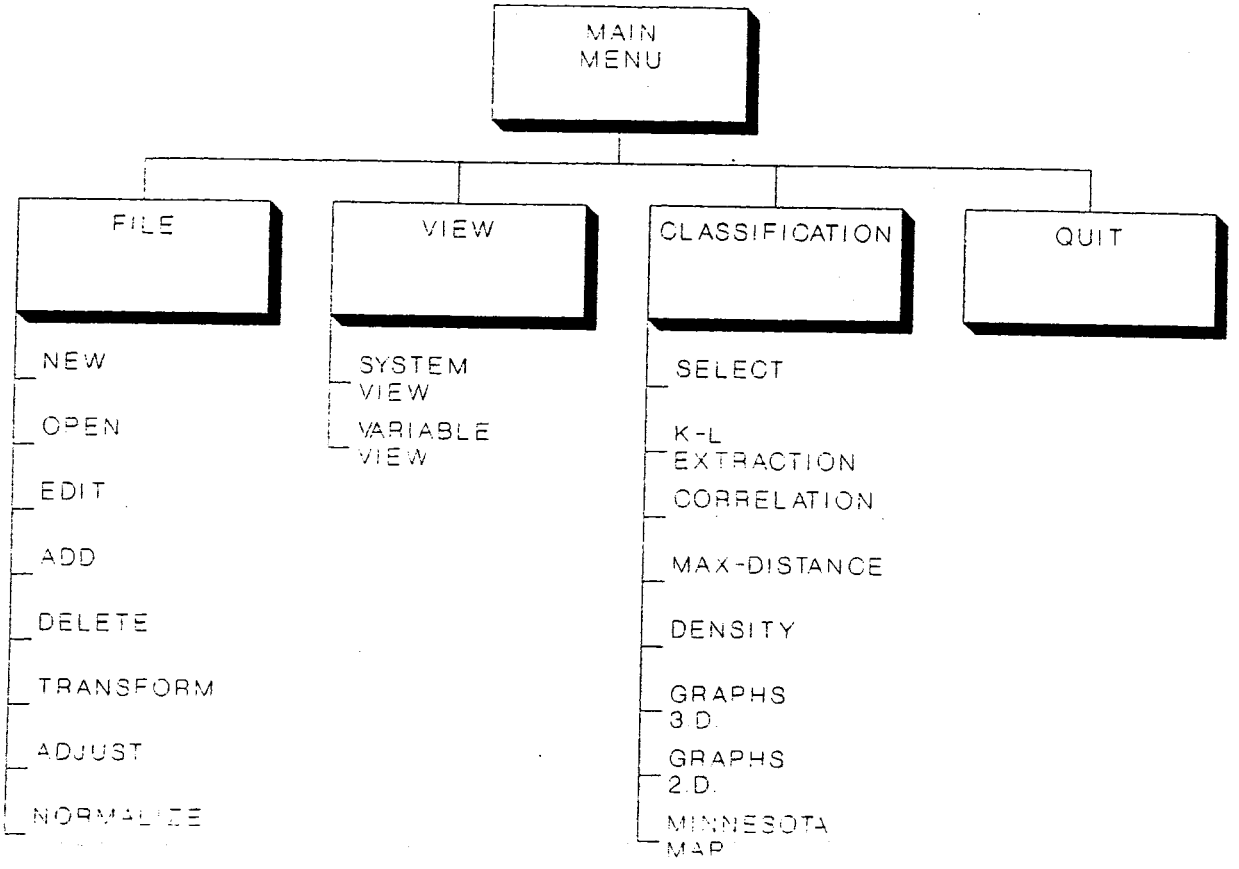

Figure B.1.1 **Structure of the classification program.**

multiplying, adding, etc. the existing indicators. So, from the miles and the cost we can have the indicator cost/mile.

♦ Scale, Normalize. By using this function the user can scale or normalize the data. In this way all the variables will receive the same weight in the classification process.

B.1.2 Display Data Functions.

There are two functions that give a graphic representation of the data base:

♦ By system. For each transit system all the data are presented on the screen and the position of the system is indicated on a map of the state of Minnesota (Fig. B.1.2).

♦ By variable. This function does not give a list with data as previous function but gives information for each variable. The maximum and the minimum value, the system distribution of the variable. The system distribution is presented also on a map of the state of Minnesota, so the user can see if there is any relationship between the variable and the location of the systems (Fig. B.1.3).

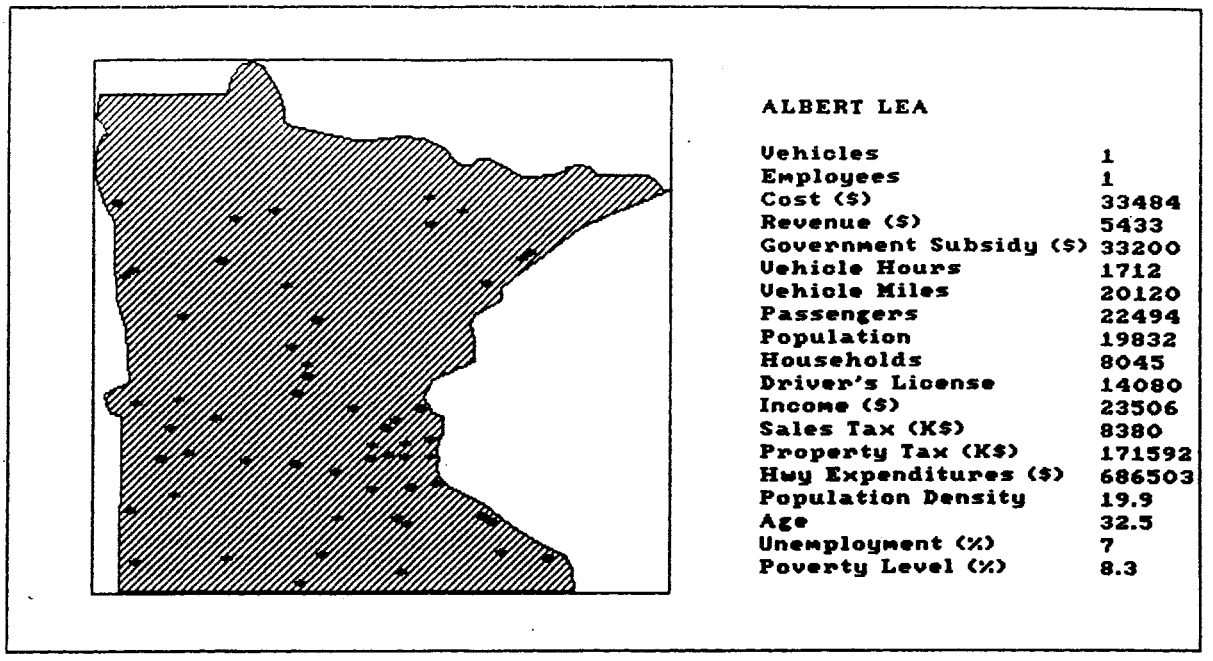

Figure B.1.2 Display data by systems.

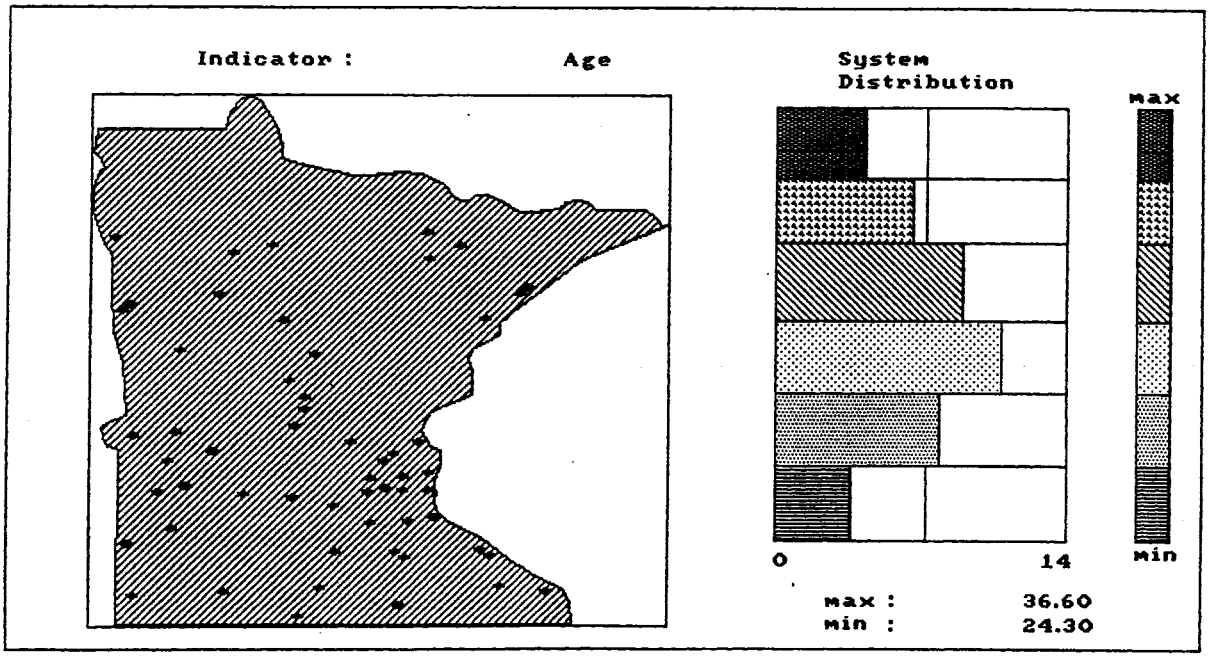

Figure B.1.3 Display data by variable.

B.1.3 Classification Functions.

After the data set is being retrieved from the data file the classification functions can be used to classify the data, to have a correlation table, or to have a graph with the classification results.

- ♦ Select (Systems, Indicators). This function gives the ability to the user to select the transit systems and the variables that will be used by the classification method or by the Karhunen-Loève feature extraction.
- ♦ Karhunen-Loève extraction. This function performs the Karhunen-Loève feature extraction method to the selected transit systems and variables. The calculated eigenvectors and eigenvalues are presented on the screen and also are stored in a results file.
- ♦ Correlation table. This function checks for multicollinearity between the variables. A correlation table is created which includes all the variables. The variables with correlation more than 90% are indicated and the user has to decide if the correlated variables will be included in the classification or not.
- ♦ Combined, Density method. These two functions execute the Combined, or the New classification

APPENDIX A

A Classification Success Index.

A method for identifying the number of clusters of points in a multidimensional Euclidean space is described and an informal indicator of the "best number" of clusters is suggested. It is a "variance ration criterion" giving some insight into the structure of the points. A familiar objective function applicable in cluster analysis is the within-group (cluster) sum of squares (WGSS). It seems natural to regard the optimal grouping of n points into k clusters as that for which WGSS is minimized. This criterion reflects a desire to find some minimum variance spherical clusters.

A.1 Methodology Used.

Suppose there are n individuals (or samples from n populations) with observations on the same v variables for each individual. We may imagine them as being represented by n points in a v-dimensional Euclidean space, P_1, P_2, ..., P_n. The variables permit the computation of an n by n distance matrix. If we denote the original v by n data matrix by $\underline{X}$, with rows given by the observed variables and with columns

given by the individuals, we can write $\underline{X} = (\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n)$, where the column $\underline{x}_i$ is a vector of the v coordinates of the point P_i. If we refer the coordinates to orthogonal axes of an ordinary Euclidean space then the distance d_{ij} between P_i and P_j will be properly defined by the function:

$$d_{ij}^2 = (x_i - x_j)'(x_i - x_j), \quad i,j = 1,2,\ldots,n \qquad (A.1.1)$$

A similar formula applies to the distance between a point and the centroid of the n points. In the approach to cluster analysis which we follow, the dispersion of a group of n points is measured by the sum of the squared distances of the points from their centroid. This sum is equal to the trace of the matrix $\underline{R}$, but may be obtained from the pairwise distances dij by applying the formula:

$$Trace\ R = n^{-1}(d_{12}^2 + d_{13}^2 + \ldots + d_{n-1,n}^2) \qquad (A.1.2)$$

If we examine a split leading to a division of the n points into k groups of $n_1, n_2, \ldots, n_k$ points ($n_1 + n_2 + \ldots + n_k = n$), then the WGSS is calculated by applying the right hand side of Eq. A.1.2 to each of the clusters separately and then summing the results. We may then write:

$$WGSS = Trace\ W = Trace\ R_1 + Trace\ R_2 + \ldots + Trace\ R_k \qquad (A.1.3)$$

where

$$Trace\ R_g = n_g^{-1}(d_{12}^2(g) + d_{13}^2 g + \ldots + d_{n_g-1,n_g}^2(g)) \quad (A.1.4)$$

with $d_{ij}(g)$ denoting the distance between points P_i and P_j in the gth cluster (g = 1, 2, ..., k). If k, the number of clusters, is not known, we proceed as follows: first we take k = 2, then k = 3, and so on. At each stage we find "the best sum of squares split", for which we calculate not only the (minimum) WGSS, but also the (maximum) BGSS and the variance ratio criterion:

$$VRC = \frac{\frac{BGSS}{k-1}}{\frac{WGSS}{n-k}} \quad (A.1.5)$$

Eq. A.1.5 give an informal indicator for the best number of groups. It is evident that this criterion is analogous to the F-statistic in univariate analysis.

When between two classification results the number of classes is the same, then one way of comparing the two results is to compare the variation within the groups to the variation between groups, by considering the matrix BW^{-1}. This matrix is the result of multiplying B, the matrix of the sum of squares between groups by the inverse of W, the matrix of the sum of squares within groups. The size of BW^{-1} indicates the general relation between the variation within groups and the

variation between groups. A large matrix is desirable as it suggests a high variation between groups compared to variation within groups. In this case, the determinant of BW^{-1} is used as an indicator of the size of this matrix. This index can not be used in cases where the two classification results do not have the same number of classes, because the determinant tends to increase as the number of classes increases.

In order to test the sensitivity of the index a small data set was used. The data set consists of nine points and there are three classes. The first class has four points, the second class has three points and the third class has two points (Fig. A.1.1). The index was tested for two different cases.

In the first case the position of one class was changing. In this way the variation within the groups was the same but the variation between the groups was changing. In that test the class A was moving towards to the class B and after a position change the index was calculated. In Fig. A.1.2 there is a line graph which shows all the values of the index. Line graph "a" describes the variation between classes A and B, and line graph "b" describes the variation between classes A and C. On the figure we can see that the variation between A and B gets the minimum value when the distance between the two classes becomes almost zero. Similarly, the variation between

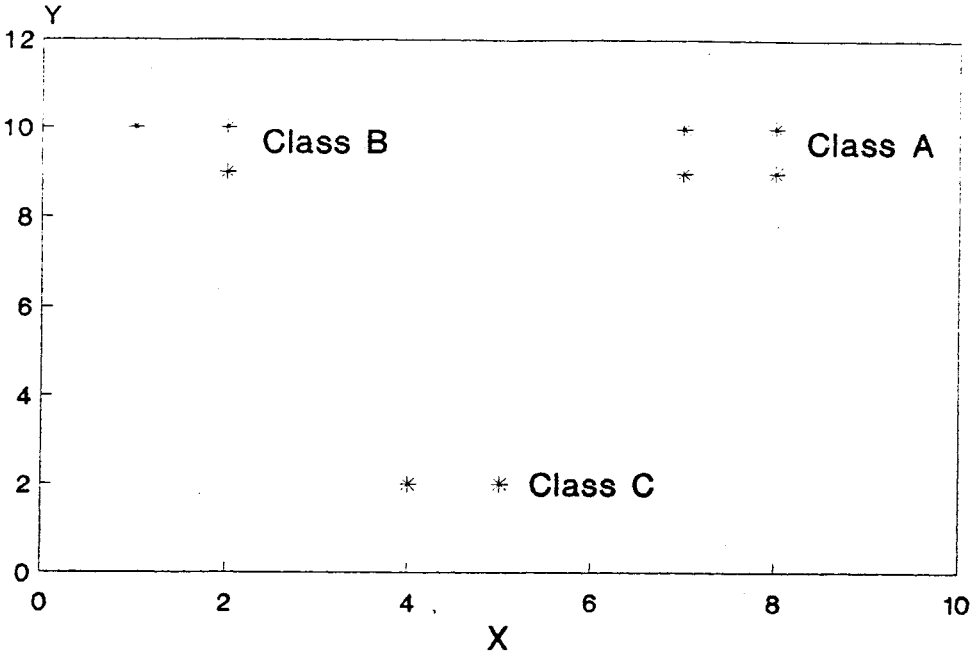

Figure A.1.1 Data set that was used for testing of the classification index.

classes A and C gets the minimum value when the distance between the two classes has the minimum value.

In the second case the positions of the cluster centers remained the same but the two points in class C become more spread out. In this way the variation between the groups was the same but the variation within the groups was changing. In Fig. A.1.3 there is a line graph which shows how the index was changing. In the beginning the index is very sensitive and for the first steps the value is reduced to the 10% of the initial value, then the differences are small.

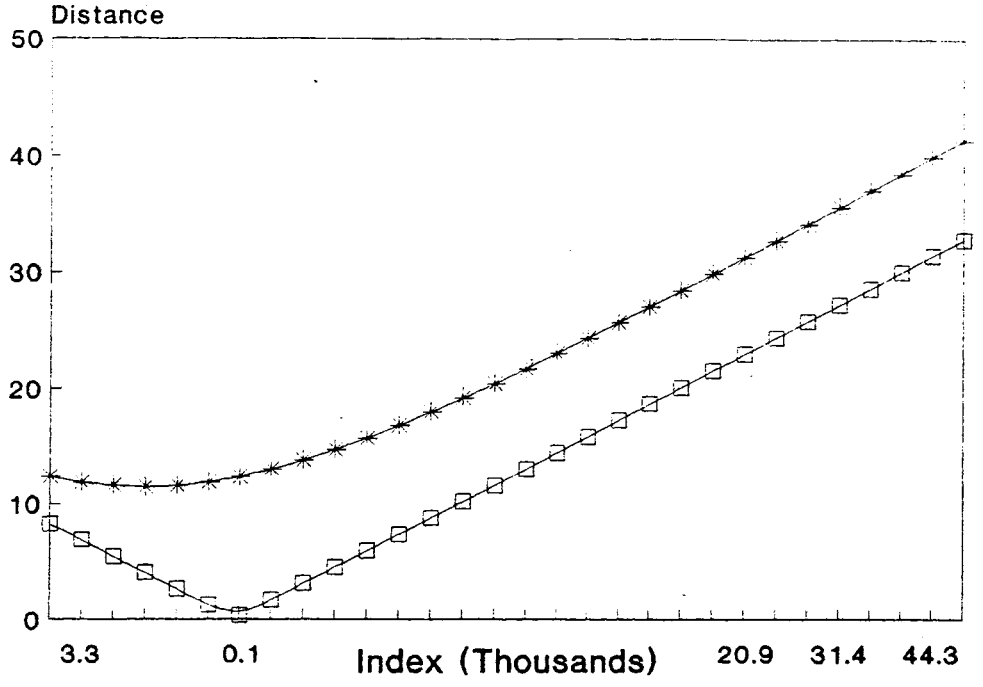

Figure A.1.2 Change of the classification index as the class A is moving.

The two testing cases show that the classification index gives good results and that is sensitive in changes of the within group variation and in changes of the between group variation. Because of the differences between the two classification indexes each index was used for a different purpose. The first classification index was used to specify the optimum number of classes and the second classification index was used in order to decide for the most appropriate classification method.

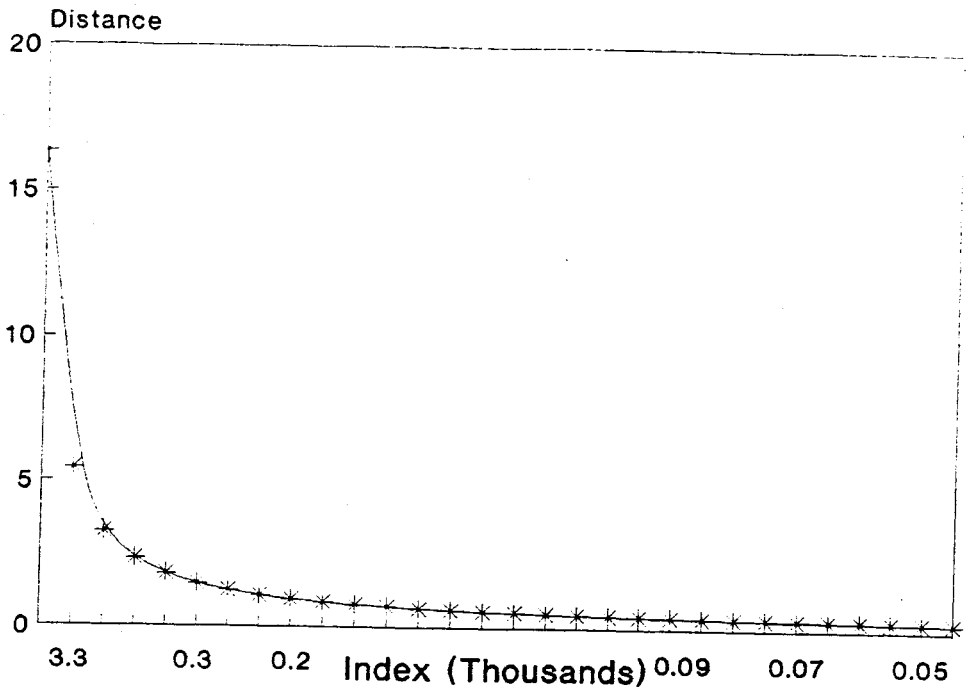

Figure A.1.3 Change of the classification index as the class C is spread out.

APPENDIX B

B Classification and Testing Programs.

In this research several classification methods were tested and many data sets were used, real and randomly generated. Also different ways to refine the data were used, the data were scaled or normalized. In order to do this work fast and more accurately two computer programs were written for these purposes. The first program was written to test the different classification algorithms by using real or randomly generated data. The second program is more complicated because it can handle data bases, apply classification methods, analyze and represent the classification results. In this section the structure of these two programs will be illustrated, and a summary of their functions will be discussed. Both programs were written in Microsoft Quick C, v5.1 and a 386 PC Compatible machine was used. Also a EGA or VGA color monitor is required to have graphic representations of the data and the classification results.

B.1 Classification Program.

This program was used to apply the classification algorithms on real data. Although data from transit system were used in

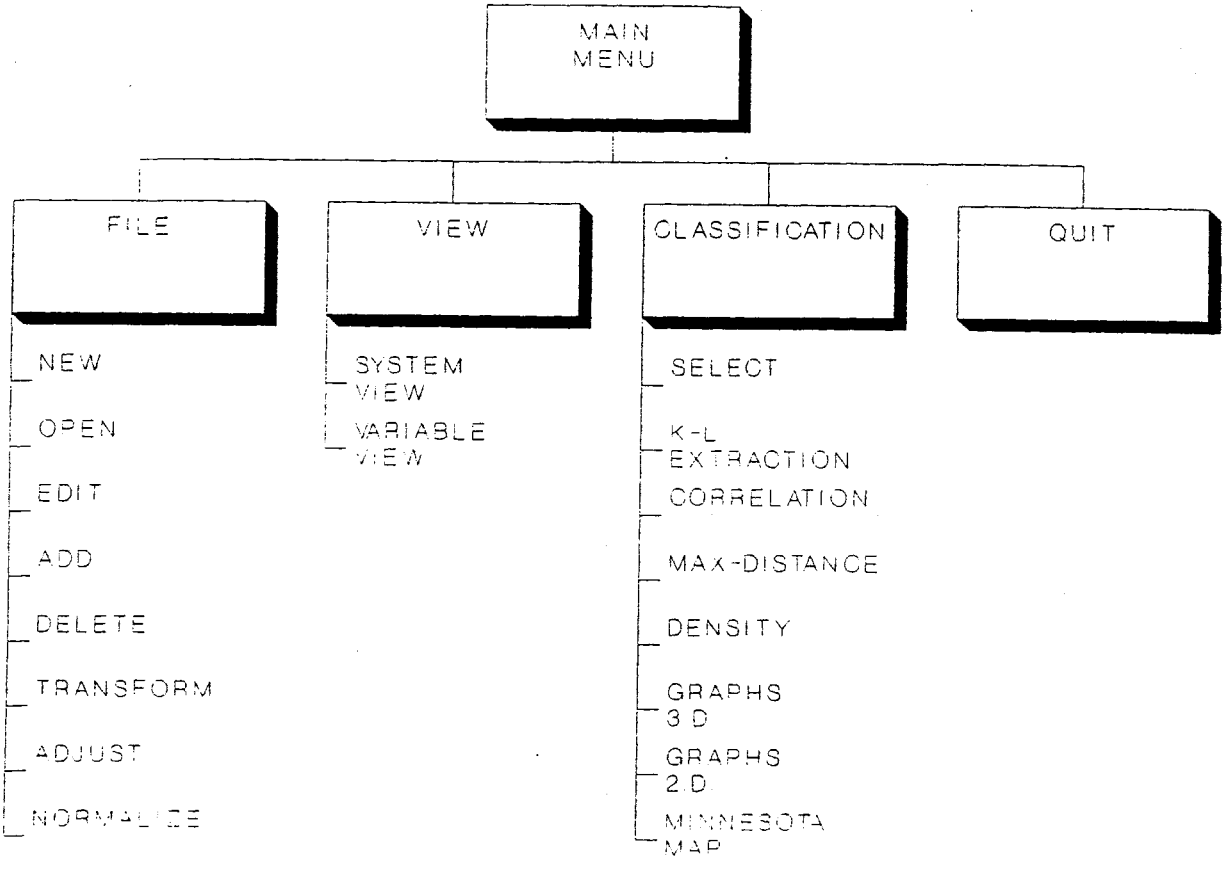

Figure B.1.1 Structure of the classification program.

multiplying, adding, etc. the existing indicators. So, from the miles and the cost we can have the indicator cost/mile.

♦ Scale, Normalize. By using this function the user can scale or normalize the data. In this way all the variables will receive the same weight in the classification process.

B.1.2 Display Data Functions.

There are two functions that give a graphic representation of the data base:

♦ By system. For each transit system all the data are presented on the screen and the position of the system is indicated on a map of the state of Minnesota (Fig. B.1.2).

♦ By variable. This function does not give a list with data as previous function but gives information for each variable. The maximum and the minimum value, the system distribution of the variable. The system distribution is presented also on a map of the state of Minnesota, so the user can see if there is any relationship between the variable and the location of the systems (Fig. B.1.3).

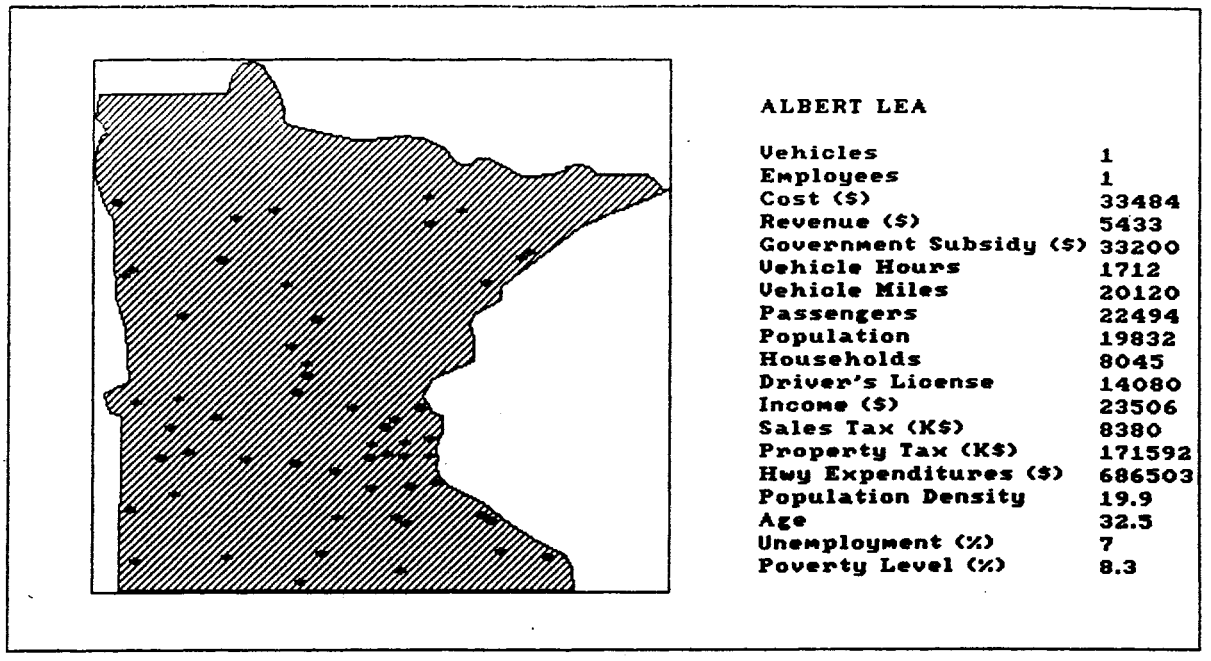

Figure B.1.2 Display data by systems.

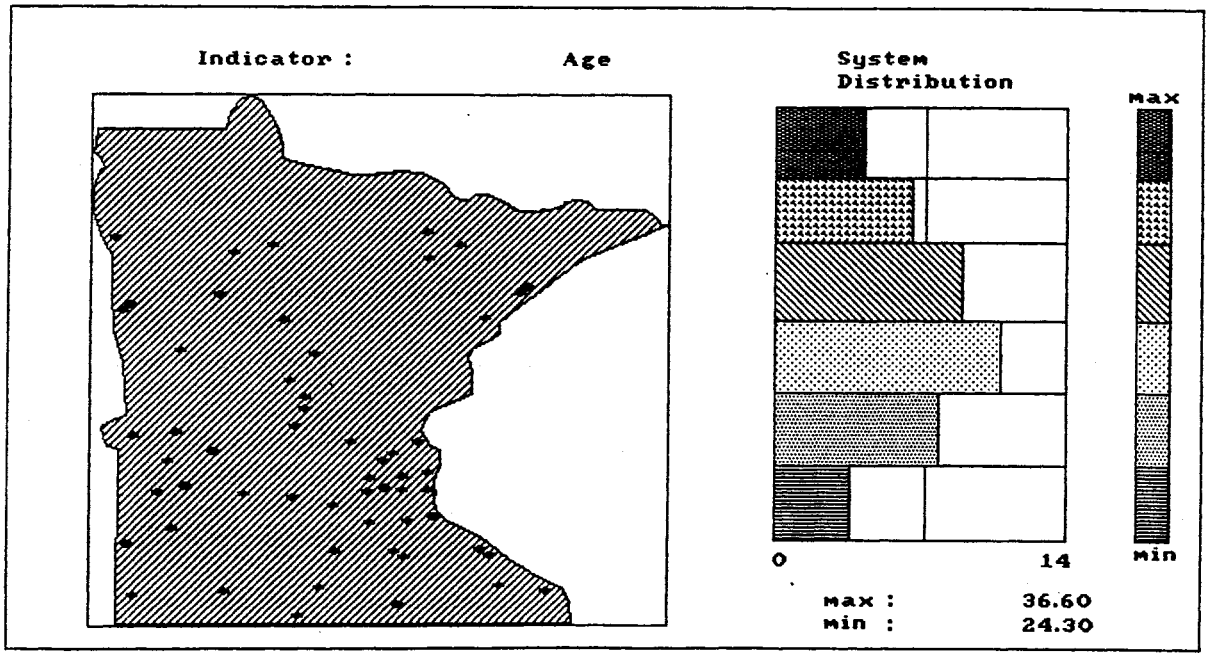

Figure B.1.3 Display data by variable.

B.1.3 Classification Functions.

After the data set is being retrieved from the data file the classification functions can be used to classify the data, to have a correlation table, or to have a graph with the classification results.

- ♦ Select (Systems, Indicators). This function gives the ability to the user to select the transit systems and the variables that will be used by the classification method or by the Karhunen-Loève feature extraction.
- ♦ Karhunen-Loève extraction. This function performs the Karhunen-Loève feature extraction method to the selected transit systems and variables. The calculated eigenvectors and eigenvalues are presented on the screen and also are stored in a results file.
- ♦ Correlation table. This function checks for multicollinearity between the variables. A correlation table is created which includes all the variables. The variables with correlation more than 90% are indicated and the user has to decide if the correlated variables will be included in the classification or not.
- ♦ Combined, Density method. These two functions execute the Combined, or the New classification

algorithm, consequently. The systems and the data that were selected are being used. A summary of the classification is represented on the screen. The classification results are stored with more details in a results file.

♦ Graphs 2.D and 3.D. These two functions present the classification results by a scatter plot. In the two/three dimensional graphs the scatter plot of two/three variables is created. In both cases the systems that are assigned in different classes are represented with different colors. In this way the user can have a view of the classes and see if there is overlapping.

♦ Minnesota map. This function, like the previous one, present the classification results by using graphics. A map with all the transit systems in Minnesota is presented, and each class of system is marked with different color. By using this function the user can observe if there are any classification patterns, e.g. systems that are located on the west part of the state of Minnesota.

B.2 Testing Program.

This program was used to test different classification algorithms by using real and randomly generated data. The program is using three windows on the screen (Fig. B.2.1):

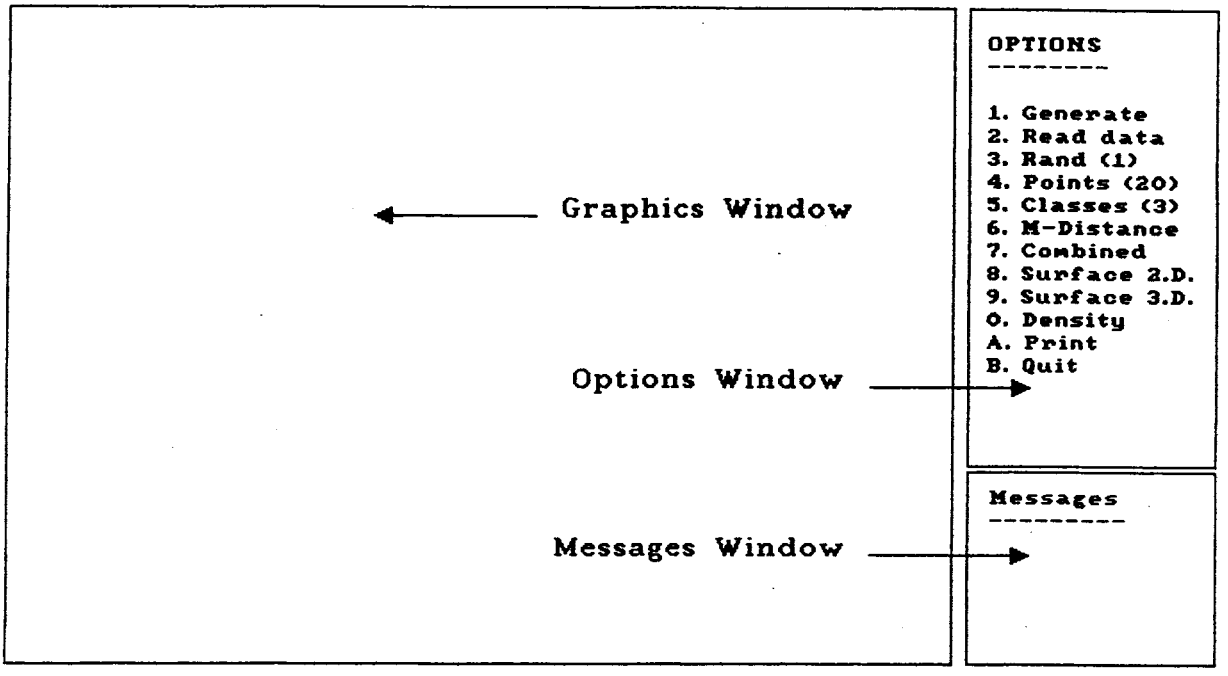

Figure B.2.1 The three windows that the testing program is using.

♦ Options Window. This window has a list with all the functions that the program can execute.

♦ Messages Window. This window shows all the messages that the gives for different reasons, e.g. in cases of errors.

♦ Graphics Window. On this window the program shows a graphical representation of the data, of the

classification results, or more complicated graphics, e.g. the cumulative potential function for a set of points.

During this research the program was updated and different functions were added or modified. The functions that are included in the latest version of the program are the following (in the same rank that appear on the Options Window):

- Generate. This function is using the random generator of the C Language, to generate a predefined number of data points. In the same time the data points are represented on the Graphics Window. In this way the user can create an unlimited number of data sets, to test the classification algorithms.
- Read data. This function allows the user to read a data set from a file. So, the method can be tested with real data, or with data that have specific characteristics, e.g. chain effect.
- Rand. This function controls the starting point of the random generator of the C Language. The default value is 1.
- Number of points. The number of points that will be generated each time that the function "Generate"

is executed. There is no limit for the number of points, but a selection of more than 40 points gives, most of the times, a "cloud" of points. the default number of points is 20.

♦ Number of classes. The number of desired classes, can be selected by this function. The maximum number of classes, that can be selected, is 9. The program has as default value 3.

♦ Max-distance, Combined, Density. These three functions execute the Maximum-Distance, the Combined, or the New classification algorithm, consequently. The most current data set, that was generated randomly or loaded from a file is being used. The results are represented on the Graphic Window by using different patterns for each class (Fig. B.2.2).

♦ Surface 2.D. This function gives a two-dimensional representation of the cumulative potential function, for a given data set. The difference values of the function are represented by the use of equipotential contours.

♦ Surface 3.D. This function works exactly like the two-dimensional function, but in this case the cumulative potential function is represented by using three-dimensional graphs.

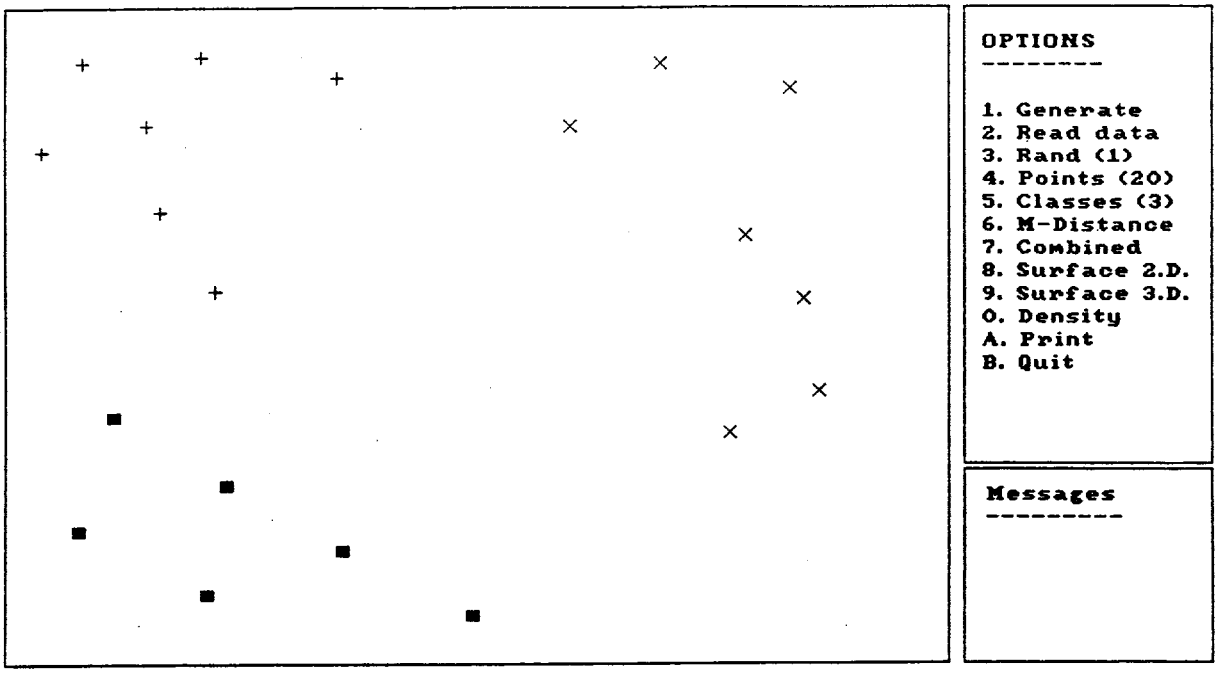

Figure B.2.2 Graphic representation of the classification results.

The structure of the program is presented on Fig. B.2.3.

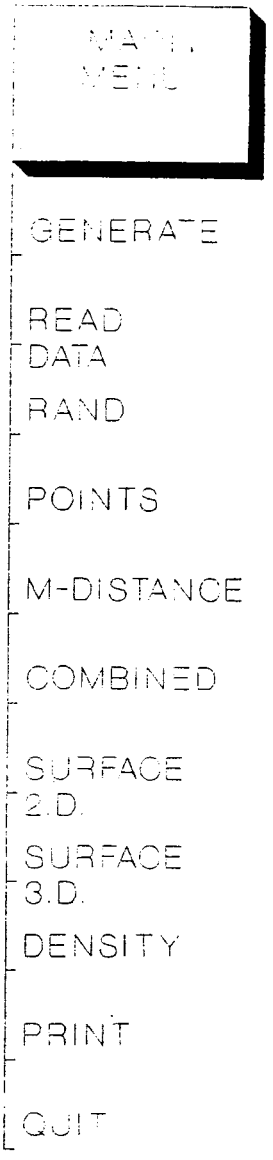

Figure B.2.3 **Structure of the testing program.**