National Library of Canada
Canadian Theses Service
Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada
Service des thèses canadiennes

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

NL-339 (r. 88/04) c

Canada

THE UNIVERSITY OF ALBERTA

FUZZY CLUSTERING AND PERCEPTIONS OF TRANSIT TRAVEL TIME COMPONENTS

by

CHARLENE LYNN ROHR

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE

DEPARTMENT OF CIVIL ENGINEERING

EDMONTON, ALBERTA

SPRING OF 1990

National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service

Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

ISBN 0-315-60297-X

NL-339 (r. 88/04) c

Canada

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR CHARLENE LYNN ROHR

TITLE OF THESIS FUZZY CLUSTERING AND PERCEPTIONS OF TRANSIT TRAVEL TIME COMPONENTS

DEGREE FOR WHICH THESIS WAS PRESENTED MASTER OF SCIENCE

YEAR THIS DEGREE GRANTED SPRING OF 1990

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(SIGNED) ..Charlene...Rohr...

PERMANENT ADDRESS:

..17.Hummingbird.Road........
..Sherwood Park..Alberta.....
..Canada.....................

DATED ..April..23.......1990

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled *Fuzzy Clustering and Perceptions of Transit Travel Time Components* submitted by Charlene Lynn Rohr in partial fulfilment of the requirements for the degree of Master of Science.

....

Supervisor

....

....

....

Date...... April 20, 1990 ......

To my sister Cheryl,
whose perseverance and dedication
is truly inspirational

ABSTRACT

This research investigated the influence of socioeconomic and mode choice characteristics on travelers' perceptions of the walk times from their homes to their public transit stops and the wait times at these stops. The imprecise nature of travelers' perceptions, resulted in the consideration of a non-traditional method of analysis - Fuzzy Cluster Analysis. This method was chosen because of its ability to analyze, mathematically, imprecise phenomena.

The basic premise of the Fuzzy Cluster Method is that membership of any object in a cluster is a matter of degree. Degrees of membership can range from zero to one, with values approaching one indicating high membership in the cluster. It is this idea of a grade of membership, versus absolute membership, that distinguishes the Fuzzy Cluster Method from classical "crisp" clustering techniques.

The objective of this research, then, was to develop clusters of perceptually homogeneous travelers with regard to public transit walk and wait times. It was hypothesized that because of the importance of perceived public transit walk and wait times in mode choice, that these clusters would represent different public transit market segments.

Up to six clusters, each with unique socioeconomic, mode choice, and perceptual characteristics were developed in this research. The properties of each cluser were explored by examining the characteristics of the cluster core. The distributions of characteristics in the cluster

cores were used to examine socioeconomic and mode choice differences between clusters. The Student t-test was employed to test whether perceptual differences between cluster cores were significant. It was concluded that differences between clusters whose perceptions were very different were significant at the 90% level of significance.

Although a few shortcomings were identified in the Fuzzy Cluster Method, it is the general conclusion of this research that this method has considerable merit as a tool for developing public transit market segments, particularly when imprecise variables such as perceptions are included in the analysis.

ACKNOWLEDGEMENTS

I wish to acknowledge the contribution of all individuals and organizations who made the completion of this thesis a reality.

The research was financially supported through scholarships from the National Sciences, Engineering, and Research Council (NSERC), Delcan Ltd., and the University of Alberta.

I thank Dr. James Bezdek for his correspondence and, particularly, for providing the source code for the Fuzzy C-Means program.

My most heartfelt thanks are expressed to Professors Doug Hunt and Stan Teply whose guidance, encouragement and continued interest contributed to the successful completion of this thesis.

Finally, my deepest gratitude to my family and friends, for without their unbounded patience and understanding none of this would have been possible.

Table of Contents

List of Tables

List of Figures

Figure Page

Figure Page

List of Abbreviations

CBD	Central Business District
CMCS	Coded Morning Commuter Survey
CU	Composite Utility
FCM	Fuzzy C-Means
LRT	Light Rail Transit
MIDAS	Michigan Interactive Data Analysis System
SSP	IBM Scientific Subroutine Package
WGSS	Within group sum of squared errors

List of Symbols

ϵ	is an element of
$\not\epsilon$	is not an element of
$\leftrightarrow$	if and only if
$\subset$	set containment
$\forall$	for all
X	Data set
n	number of objects
x_k	k^{th} object
R^p	real p-dimensional space
R^+	the real interval $[0,\infty)$
Q_c	Crisp c-partition space
Q_f	Fuzzy c-partition space
c	number of clusters
V_{cn}	set of all real possible c x n cluster membership matrices
U	one possible membership matrix solution $(U \epsilon V_{cn})$
μ_{ik}	membership of the k^{th} object in the i^{th} cluster
v_i	centroid of cluster_i
d_{kj}	measure of dissimilarity between any two points x_k and x_j
s	exponential weight
M	dissimilarity metric
F(U;c)	Partition Coefficient
H(U;c)	Partition Entropy Measure
$HCE(U)_i$	Cluster Entropy Measure

1. INTRODUCTION

1.1 Problem Statement

Marketing in a modern public transit organization is much more than advertising and promotion; it is an approach to transit management effecting the analysis, planning and implementation of transit services. The 1985 Canadian Transit Handbook recognizes the importance of public transit marketing stating "ultimately, the success or failure of any organization depends on how well it recognizes its market, tailors its products to that market, promotes its product, and then delivers the goods to the satisfaction of its customers" (CUTA, 1985, p.22-2).

Market segmentation is a means of distinguishing different consumer groups in a market. Market segments are typically defined according to aspects of consumer behaviour, socioeconomic characteristics, or other attributes relevant to the particular segmentation problem (Hensher, 1976).

In regards to mode choice, and the competition between public transit and automobile, two key public transit attributes which effect whether individuals will choose public transit or not are the walk times from their homes to their transit stops and their wait times at these stops. Truly, however, it is not the actual values of these variables which influence mode choice but rather the travelers' perceptions of these values. This implies that to attract and to keep transit riders transit agencies must not only minimize these travel time values, but they must also ensure that travelers' perceptions of these times reflect true values.

Because of the importance of travelers' perceptions of the walk times from their homes to their transit stops and the wait times at these stops, in mode choice, these variables were used for segmentation of the public transportation market in this research. The resulting market segments are therefore defined according to the accuracy of the travelers' perceptions. The socioeconomic and usual mode choice characteristics of the travelers were also included as segmentation variables.

Based on these results, transit market strategies particular to specific perceptual and socioeconomic groups can be developed. For instance, for market segments where travelers' perceptions of public transit travel times differ substantially from actual travel time values, marketing strategies aimed at altering these travelers' perceptions can be introduced. On this subject Gilbert and Foerster (1977) state, "Just as consumers' purchases of automobiles, for example, are influenced by the intense marketing efforts made by automobile manufacturers, so too can transit use be influenced by improving the travelers' perceptions to and attitudes about transit" (Gilbert and Foerster, 1977, p. 322).

For market segments in which travelers' perceptions accurately reflect reality, marketing strategies concentrating on public transit comfort, convenience or cost might be a more advantageous means of attracting public transit passengers.

The mathematical technique employed to group, or to cluster, the individuals into the market segments is called the Fuzzy Cluster Method. Fuzzy Cluster analysis is particularly suitable for this analysis because it allows, mathematically, for imprecision in the process being analysed and human perceptions are imprecise by nature. The basic premise of Fuzzy Cluster theory is that membership of any object (in this case individual) in a cluster is a matter of degree. The degree of membership is a real number between zero and one, with values approaching one indicating high membership into a cluster. It is this idea of grade of membership, versus absolute membership, that distinguishes Fuzzy Cluster theory from classical "crisp" cluster theory.

The results from this research also provide interesting insight into how different groups of travelers perceive public transit service attributes. This is important for a better understanding of how traveler mode choices are made, and has implications for better travel demand prediction. At present, "usual" behavioural mode choice models, i.e. logit models, are based on objectively measured values of relevant mode service characteristics. These are, for example, travel time values, wait time values, travel costs etc. - values travelers encounter in reality. However, in choice behaviour theory, choice of items such as travel mode are based on travelers' perceptions of the characteristics of the various modes (Clark, 1982). Therefore, it is logical to argue that when modeling mode choice behaviour perceived values, not objectively measured values, should be used. Unfortunately, travelers' perceptions of modal attributes are usually not available, and at this time are impossible to predict. According to Stopher and Meyburg (1975), "this appears to be the only basis on which measured (objective) values can be justified" (Stopher and Meyburg, 1975, p. 37). That is, because of the relative ease of prediction and collection of measured data, they are the variables generally used in mode choice behaviour models.

The development of a travel time perception model would solve this problem. Theoretically, this model could be used to convert actual measured travel times into perceived travel times prior to use in a behavioural choice model. Development of such a model is well beyond the scope of this research; however, results from this research investigating travelers' perceptions and the factors that affect these perceptions may be viewed as a first step in the model development.

1.2 Research Objective

The primary objective of this research is to use travelers' perceptions of public transport walk and wait times, socioeconomic characteristics, and usual mode choice to develop public transportation market segments. In developing these market segments two secondary objectives are defined. These are:

1. To introduce Fuzzy Set Theory to transportation engineering practice in Edmonton, Canada.
2. To better understand travelers' perceptions of time-related public transit attributes for improved prediction capabilities.

1.3 Scope

In 1983 the Transportation Department of the City of Edmonton initiated a Downtown Commuter Survey to investigate morning peak hour home-to-work travel behaviour. In this survey over 1700 individuals who worked in the Edmonton Central Business District (CBD) were asked questions regarding: (1) their socio-economic character, (2) the trip they made, and (3) alternatives to the trip they made. The survey was called the Morning Commuter Survey. For detailed information on this survey see Hunt (1984).

The information required for the present research was primarily derived from this "1983 Morning Commuter Survey" database. For this reason the research is limited to those issues included in the survey, particularly to the investigation of travelers' home-to-work trips to the Edmonton CBD.

1.4 Organization of the Thesis

Chapter 2 reviews travel perception literature and transportation segmentation research projects relevant to this study.

Chapter 3 describes Fuzzy Cluster theory and provides justification for its use in this research.

Chapter 4 describes the variables used in the Fuzzy Cluster analysis procedure.

Chapter 5 presents the Fuzzy Cluster results.

Chapter 6 examines the characteristics of the "best" cluster structures identified in Chapter 5.

Chapter 7 documents the development of two multivariate linear regression models. The results from these "more traditional" models are compared to those obtained from the Fuzzy Cluster Method.

Chapter 8 summarizes the main findings of this research, evaluates the research procedure, and provides a practical example of how the results of the cluster analysis can be used for development of public transit market strategies.

2. LITERATURE REVIEW

The primary objective of this research is to use travelers' perceptions of public transport walk and wait times, socioeconomic characteristics, and usual mode choice to develop public transportation market segments. This objective encompasses, essentially, two areas of transportation research: travel time perception research and market segmentation analysis.

Previous travel time perception research relevant to this research describes: (1) the problems associated with the use of perceived and objective data and (2) studies which address the question of whether travelers' perceptions of travel times differ from true measured values. A review of this literature is presented in the first part of this chapter.

The second part of this chapter describes a number of market segmentation research projects which focus specifically on transportation issues.

2.1 Travel Time Perception

Clark (1982) states that travelers' perceptions of travel times and travel costs may be represented either by (1) actual measured values of these variables or (2) values estimated by the individuals in travel surveys. In comparison to true perceptions, both are subject to inaccuracies; however, he believes that the values estimated by travelers in survey situations are the better attempt of translating perceptions into measurable quantities. Following is a brief description of the biases present in both methods.

2.1.1 Stated Value Biases

Values reported by travelers in survey situations are imperfect approximations of their perceptions because they are subject to the following biases:

1. Ex-post rationalization

   This reporting bias arises from travelers, either consciously or subconsciously, distorting their perceptions to appear more logical. In general it is not possible to say anything about the magnitude of this bias, as true perceived values, at present, are unable to be determined.

2. Rounding Bias

   Rounding biases occur because of the difficulty people have in distinguishing small units of time. Watson (1971) hypothesizes that people are more likely to round down reported times for their chosen mode, and round up for the rejected mode. Again, this may be partly due to a rationalization of their decision to use their chosen mode.

   An example of the magnitude of this bias is seen in research by O'Farrell and Markham (1974), who reported commuters tended to round off in-vehicle times to five minute values.

3. Lack of information

    Some travelers may be unaware of the travel times and costs of their alternatives. In this case reported values may not represent perceived values, but may simply be guesses.

    Hensher (1975) hypothesizes that there are two periods in the mode choice decision: the habit period and the decision period. He argues that during the habit period individuals are not seeking information on alternative modes of travel and therefore will continue to use the same mode of travel until some "strong stimulus" forces them to re-evaluate their present mode choice. At this time travelers may seek out information on alternative modes. He argues that since people are not constantly changing their travel patterns, the majority of commuters are in the habit period, and therefore, any estimates of attributes of alternative modes are simply "uninformed guesses".

In this research, travelers' perceptions of public transit walk times and wait times are represented by values obtained from survey responses. The aforementioned biases, however, are acknowledged in the analysis of the results.

2.1.2 Measured Value Biases

Actual measured values are also subject to bias. These values are usually estimated from average speeds, assumed routes, operating schedules, etc. and therefore represent average journey times rather than the times that may be actually experienced by the individual (Watson, 1971).

In this research actual public transit walk times and wait times are estimated from measured walk distances, assumed walking speeds, and assumed public transit service characteristics. For further details see Section 4.3.1.

2.2 Results of Past Research Investigating Perceived and Measured Modal Attribute Data

Past research addressing the question of whether commuters' perceptions of modal attributes differ significantly from measured values has resulted in conflicting findings. Research by Quarmby (1967), O'Farrell and Markham (1974), Heggie (1976), and Meyburg and Brog (1981) suggests that measured and perceived values are different, by different amounts, for transit and car users. Watson (1971) and Algers et al (1975) suggest that in some cases this may not be true. Following is a summary of results obtained by these individuals.

Citing evidence presented by Lisco (1967), Watson (1971) states that the differences between perceived and actual journey times are approximately normally distributed with an expected value of zero. This suggests that there is no difference, or consistent bias, between perceived and actual journey times.

In Sweden, Algers et al (1975) investigated the relationship between perceived travel times and estimated actual travel times for car and transit users. Based on findings in Sweden (see Figure 2.1), they concluded that both mode users "perceive the same variable values (travel time values) as those measured by traffic engineers" (Algers et al, 1975, p. 40). The sample sizes they used to establish this fact, however, were quite small (approximately 35 users).

Quarmby (1967), in a comparison of bus users' and car users' estimations of bus travel times, found car users' perceptions to be approximately 20% higher than bus users'. He estimated that half of this difference was attributable to an actual difference in walking and waiting times (i.e. bus users being more familiar with schedules and therefore actually waiting less time) and half to a genuine difference in the perceptions of car and bus users. He gave no basis for this estimation. As well, he was unable to determine the accuracy of either of these groups' perceptions as he did not have any estimated actual travel time values. With regard to perceptions of car travel time, he found no significant difference between the car and bus users.

O'Farrell and Markham (1974) investigated the extent to which commuters' perceptions of public transport wait and travel times differed from actual times and whether the relationship between perceived and actual values was dependent on mode user group. They investigated the

FIGURE OMITTED

DUE TO

COPYRIGHT

RESTRICTION

Figure 2.1 Perceived Versus Estimated Travel Times for Automobile and Public Transit Users (Algers et al, 1975, p. 41)

perceptions of three user groups: train users, bus users and car drivers. The results for the bus users and car drivers are shown below.

Travel Times	Users	Time		Absolute Difference (min)	Relative Distortion %
		Actual (min)	Perceived (min)		
Perceived In-Vehicle Bus Times to Work	Car	20.31	24.22	3.91	25.86
	Bus	21.79	25.15	3.35	16.87
Perceived In-Vehicle Bus Times from Work	Car	23.41	29.03	5.62	31.93
	Bus	24.53	28.17	3.64	17.55
Perceived Bus Wait Times, Journey to Work	Car	4.02	7.50	3.49	116.76
	Bus	3.93	5.68	1.76	75.59
Perceived Bus Wait Times, Journey from Work	Car	5.85	17.25	11.40	245.28
	Bus	5.13	13.90	8.77	204.90

These results indicate that both the car and bus users overestimated the bus in-vehicle and bus wait times. In all four categories, the average car drivers' overestimations were larger than the bus users'. It is interesting to note that both the car users' and bus users' overestimations of evening wait times were much greater than those for the morning trip.

Based on these results, particularly those for the bus wait time perceptions, O'Farrell and Markham concluded that the use of objective public transport data in urban planning models needed to be questioned, since actual travel times seldom reflected the subjective images of the commuters.

Heggie (1976) examined car users' perceptions of the characteristics of the bus mode. He divided the work journey into 5 time segments: walk home to bus, wait, ride in bus, interchange time, walk bus to home. He found that within each segment, on average, bus travel times reported by car users were substantially higher than those reported by bus users. He states that this "difference cannot simply be explained by a lack of knowledge. If car users simply know less about bus travel, one would expect to find some error in their responses, but only a random error without this substantial and consistent bias" (Heggie, 1976, p. 21). Whether the bus users' perceived travel times differed from the actual travel times was not investigated.

More recently Meyburg and Brog (1981) compared reported and estimated actual travel times by automobile and transit users in Munich, West Germany. Reported travel times were for all trip purposes. Their results are summarized below.

	Avg. Mis-estimation of travel time by transit (%)	Avg. Mis-estimation of travel time by car (%)
Transit Riders	+10.4	+4.0
Automobile Drivers	+28.9	+8.4

In this study sixty-one percent of the automobile drivers overestimated their transit travel time by more than 20%.

2.3 Summary Findings of Past Research

Previous research strongly suggests:

1. That there is, in fact, differences between those values actually measured, and those perceived by

travelers.

2. That automobile drivers tend to have much larger overestimations of public transit travel time than do public transit users.

For the most part, these studies have examined travel time perception differences between different mode user groups. Investigations of other variables which might define, or categorize, groups with similar perceptions have not been made. It is possible, then, that the mode defined groups of this previous research might be comprised of smaller, more perceptually homogeneous, groups. For instance, male car drivers' estimations of public transit travel times might be more accurate than female drivers'; or older car drivers' more accurate than younger car drivers'. If this is true, average travel time estimations for car drivers, as a group, do not truly represent the perceptions of any of the individuals in this group.

The analysis approach of this research investigates the basis on which individuals might be perceptually similar by using cluster analysis, or Fuzzy Cluster Analysis, to determine perceptually similar groups. With this approach, no *a priori*, potentially biasing, assumptions regarding the characteristics of these groups are made. Whether perceptually similar groups are defined by their usual mode choice, or by their gender or age is determined by the nature of the cluster process itself.

2.4 Results of Past Market Segmentation Research

With regard to public transit, past market segmentation research has concentrated on segmenting the population according to surveyed consumer attitudes towards transit. Examples of two such studies are by Nicolaidis and Dobson (1975) and Recker and Golob (1976).

Nicolaidis and Dobson (1975) used segmentation to determine population segments which had similar public transit service preferences and priorities. They based their segmentation on "attitudinal ratings" determined from mode-independent judgements of the importance of specific transit characteristics. Those travelers who regarded the importance of attributes in a similar manner formed homogeneous groups which were then "cross-classified" with various socioeconomic variables and activity patterns to determine if there was any link between these variables and the defined segments. The socioeconomic variables race, education, and age were found to be strongly related to the perceptual groupings of the travelers.

On the applicability of market segmentation for their analysis Nicolaidis and Dobson concluded:

> "The fundamental psychological tenet around which this report centers is that people have different preferences but no individual, at least no representative individual, is totally distinct. This sharing of common preference patterns allows the understanding of how alternative innovative urban transport designs variously benefit different population segments." (Nicolaidis and Dobson, 1975, p. 294)

Recker and Golob (1976) hypothesized that individuals' attitudes towards public transit choice alternatives were a function of the public transit supply, and that these attitudes were reflected in mode choice behaviour. In their research they used market segmentation to obtain groups of travelers with similar public transit choice constraints. The resulting segments were labelled: the "mobile", the "inappropriate bus routing", the "poor bus accessibility", the "carless", and the "busless". Factor analysis was used to determine latent perception dimensions describing the respondents' perceived satisfaction with specific work-trip attributes for each segment. In brief, latent factors are linear combinations of variables that account for as much of the total variation in the data, with as few factors as possible. In Recker and Golob's research, the specific work-trip attributes considered in the factor analysis procedure included comfort, vehicle safety, privacy, and low riding time. Logit mode choice models, based solely on subsets of descriptive attribute ratings chosen to represent the latent perception factors were then developed. These were concluded to "provide useful diagnostic information on which of the many attributes describing modal alternatives are determinant in choice" (Recker and Golob, 1976, p. 309).

Stopher (1977) developed market segments for the destination choice of non-grocery shopping locations. He hypothesized that persons from a given socioeconomic group would be more likely to have homogeneous perceptions of what make shopping locations attractive. He developed perceptual spaces for several socioeconomic groups based on travelers' preferences for various attributes of the shopping locations. He concluded that length of residence and age were reasonably powerful market segmentation variables.

3. DATA ANALYSIS TECHNIQUE

This chapter describes the theory of Fuzzy Cluster Analysis. To best set forth the many concepts of this theory, the chapter proceeds as follows. First a description of general cluster analysis theory and justification for its use in this research are provided. Next, the concept of uncertainty and its relevance to this research is discussed. Finally, the details of the Fuzzy Cluster Method are presented.

3.1 Cluster Analysis

Cluster analysis is the "partitioning of a collection of objects into disjoint subsets of clusters" (Windham, 1983, p. 271). Objects which belong to the same cluster have common properties which distinguish them from the members of the other clusters. The basic assumption for the use of this analysis is that some underlying pattern exists in the data.

There are three different clustering methods: hierarchical methods, graph-theoretic methods, and objective function methods.

The hierarchical methods either gradually agglomerate all objects until eventually only one cluster is defined; or oppositely, start the clustering process with all objects belonging to a single cluster and then split this cluster until all that is left are single objects. This merging or splitting process is based on a clustering criterion which defines similarity between points.

Graph-theoretic methods require the development of node-graphs to represent the data set. Connectivity of nodes is dependent on some defined measure of node similarity.

The basis of the objective function method is an objective function which explicitly measures the "desirability" of clustering individual points. Thus, by either maximizing or minimizing this function, optimal cluster structure is determined. According to Bezdek (1981), mathematically speaking, this method is considered to be "the most precise formulation of the clustering criterion" (Bezdek, 1981, p. 47).

3.1.1 Justification for Cluster Analysis in this Research

Cluster analysis is considered a suitable statistical analysis technique for this research because:

1. As was discussed in Section 2.3, a downfall of past travel-time-perception research was the assumption that groups of individuals defined by regular mode choice are perceptually homogenous. Using cluster analysis to segment the population requires no assumptions regarding group structure. Rather, what it does is group together travelers who have similar perceptions of public transport walk and wait times. These defined groups may then be examined for socioeconomic and travel characteristic trends.
2. The resulting clusters of perceptually similar individuals may be selected as "target" markets for

public transit planning, service scheduling, and design.

3. Classification, or clustering, may be considered to be an intermediate step in what would be viewed as the ultimate objective of being able to model travelers' perceptions.

3.2 Imprecise Data

For investigations of physical processes, the choice of mathematical model type depends on many factors. One of these factors is the amount and source of uncertainty in the physical process. Bezdek (1981) identifies three sources of uncertainty that may be present. There may be uncertainty due to:

1. Inaccurate measurements
2. Random occurences
3. Vague descriptions

He suggests that each is most adequately described by deterministic, probablistic, and fuzzy models, respectively.

In this research, travelers' perceptions of public transit time components are investigated. Assuming that perceptual trends exist amongst different groups of travelers (Section 2.3) implies that these perceptions are not random; nor are they "uncertain" simply because of inaccurate measurement. These variables are, by nature, inexact.

This "inexactness" problem is described by Esogue (1986) in his discussion of the problems of modeling the human decision process. He states that because the human decision process is a very complicated process, one which is influenced by such qualitative factors as "emotion, perception, shifting and imprecise knowledge states etc.", it impossible to represent mathematically with absolute certainty (Esogue, 1986, p. 283). Since perceptions are an important component in this process one would hypothesize that they too are impossible to represent mathematically, with absolute certainty.

It is reasonable, then, to suggest that if a system is uncertain or imprecise in nature it should be analyzed by a mathematical technique which allows for imprecision. The clustering techniques described earlier do not provide for any imprecision, an object either belongs to a cluster or it does not.

Problems of this nature led to the development of Fuzzy Set Theory. This theory "provides a natural way of dealing with problems in which the source of imprecision is the absence of sharply defined criteria of class membership rather than the presence of random variables" (Zimmerman, 1985). It provides a mathematical framework in which imprecise phenomena can be studied.

Clustering, or categorizing travelers according to their travel time perceptions is subject to many sources of "uncertainty". One source is in the individuals' memberships into the clusters. One would not expect travelers to divide into "exact" categories but rather would expect that some relatively stable, yet "inexact" patterns to exist. For instance, one would expect regular public transit users to have more accurate perceptions of public transit walking and waiting times than non-public transit users (Chapter 2).

Another source of imprecision exists in the definition of the clusters. In the final results, clusters are defined by the degree of accuracy of the travelers' perceptions. Potentially there could be both "semi-crisp clusters" defining groups of travelers whose perceptions are exact, or oppositely, whose perceptions are "way-out", and in-between "fuzzier clusters" defining groups of travelers who are tending towards either these exact or way-out perceptions.

For this research, because of the imprecise nature of travelers' perceptions, it is concluded that Fuzzy Cluster analysis is a suitable method of analysis for the segmentation problem.

3.3 Fuzzy Cluster Analysis Theory

The discussion that follows is, unless otherwise specified, derived primarily from Bezdek (1981).

3.3.1 Clustering Problem Definition

Bezdek (1981) defines the clustering problem as follows:

Let $X=\{x_1,...,x_n\}\subset R^P$ be a subset of n items x_k, of the real p-dimensional space R^p. Let c, the number of clusters, be an integer such that $2 \leq c < n$; and let V_{cn} denote the set of all real possible c x n matrices where the i,kth entry, μ_{ik}, is the membership of the kth object in the ith cluster.

The basic objective of the clustering technique is to divide these n objects, where each object is characterized by a value in each of p dimensions, into c clusters. The number of clusters, c, is generally not known in advance (but must be specified for a given calculation). Cluster centres $v=(v_1,v_2,....,v_c)$ are also characterized by a value in each of p dimensions ($v_i \epsilon R^p$ for $1 \leq i \leq c$), and define the centroid of the cluster.

The partitioning of all elements in X into c clusters will be described by a cxn matrix, $U \epsilon V_{cn}$, which contains the membership function values for each object x into each cluster c. The diagram below illustrates that the i,kth entry of U, μ_{ik}, indicates the membership of the kth object in the ith cluster.

$$U = \left[\begin{array}{c|ccccccc} & x_1 & x_2 & x_3 & \cdots & x_k & \cdots & x_n \\ \hline c_1 & & & & & & & \\ c_2 & & & & & & & \\ \vdots & & & & & & & \\ c_i & & & & & u_{ik} & & \\ \vdots & & & & & & & \\ c_c & & & & & & & \end{array}\right]$$

The first step in defining the clustering problem is to determine a suitable measure of dissimilarity between points. The measure of dissimilarity, d_{kj}, between any two points x_k and x_j must satisfy the following criteria:

1. $d(x_k,x_j) = d_{kj} \geq 0$
2. $d_{kj} = 0 \leftrightarrow x_k = x_j$
3. $d_{kj} = d_{jk}$

Because of these properties, the (X x X) matrix d, will be positive definite and symmetric.

$$d:[X x X] \rightarrow R^+$$

For the clustering problem let d_{ik} be a generalized measure of separation between the point x_k and the cluster centre v_i, for some arbitrary p x p matrix M, which is positive definite and symmetric, where

$$d_{ik} = \|x_k - v_i\|_M = [(x_k - v_i)^T M (x_k - v_i)]^{1/2}$$

If M is chosen as the identity matrix, $M=[I]_{pxp}$, then d_{ik} is simply the distance between x_k and v_i in p-space. The importance of this arbitrary matrix, M, will be discussed later in this chapter (Zimmerman, 1985).

3.3.2 Set Membership

In classical (crisp) set theory a single element, x, either does or does not belong to a set A. One can define a function, u(x), which defines set membership, in which 1 indicates membership and 0 non-membership. That is:

$$\mu_A(x) = \begin{array}{l} 1,\ x \epsilon A \\ 0,\ x \not\epsilon A \end{array}$$

Contrary to this is the basic premise of Fuzzy Set Theory which is that membership of any element into a set A is a matter of degree. In this theory the membership function measures the degree of membership for the element x in set A. The degree of membership is always a real number between zero and one, with values approaching one indicating high membership in the set.

3.3.3 The Crisp Clustering Algorithm

Crisp clustering algorithms assign each object x to exactly one cluster. Bezdek (1981) defines the following criteria which must be met if the matrix $U=[\mu_{ik}]\epsilon V_{cn}$ is to represent a solution to the crisp c partitioning of X:

1. All elements in the U matrix must be either one or zero, depending on whether x_k is in the i^{th} cluster or not.

   $\mu_{ik} \epsilon \{0,1\} \qquad 1 \leq i \leq c \qquad 1 \leq k \leq n$

2. Each element x_k belongs to exactly one of the c clusters.

   $\sum_{i=1}^{c}\mu_{ik} = 1 \qquad 1 \le k \le n$

3. No cluster is empty nor contains all elements of X.

   $0 < \sum_{k=1}^{n}\mu_{ik} < n \qquad 1 \le i \le c$

The last two criteria ensure that there is at least one "1" in each row of the U matrix and only one "1" in each column for a given x. For example, if 4 objects $X=\{x_1,x_2,x_3,x_4\}$ are to be clustered into 3 partitions, the following U matrix would be one possible solution:

$$U = \begin{array}{c|cccc|} & X_1 & X_2 & X_3 & X_4 \\ \hline C_1 & 1 & 0 & 0 & 0 \\ C_2 & 0 & 0 & 0 & 1 \\ C_3 & 0 & 1 & 1 & 0 \end{array}$$

The set of all matrices which then define a crisp c-partition space for X are defined as:

$$Q_c = \{U \epsilon V_{cn} | \mu_{ik} \epsilon \{0,1\} \forall i; \sum_{i=1}^{c}\mu_{ik}=1\forall k; 0< \sum_{k=1}^{n}\mu_{ik}<n\forall i\}$$

The most extensively used clustering criterion is the classical within group sum of squared errors (WGSS) objective function defined as:

$$J(U,v) = \sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}f(d_{ik})$$

where:

$$f(d_{ik}) = d_{ik}^{2} = \|x_k - v_i\|^2 = (x_k - v_i)^T(x_k - v_i)$$

$U = [\mu_{ik}]\epsilon Q_c$ is crisp

$v = (v_1, \ldots v_c),\ v_i \epsilon R^p$

The cluster centres are defined as the centroids of the clusters and are calculated by the following formula:

$$v_i = \frac{\sum_{k=1}^{n} \mu_{ik} x_k}{\sum_{k=1}^{n} \mu_{ik}} \qquad \text{for each dimension in } R^p \qquad (1)$$

Since $\mu_{ik}=1$ if and only if x_k is in the i th cluster, the objective function, J(U,v), is simply a measure of dissimilarity between points in a particular cluster and the cluster centre. More specifically, since the dissimilarity term, d_{ik}, is defined as the Euclidean distance between x_k and v_i (that is, $M=[I]_{pxp}$) and this distance term is squared, the objective function J(U,v) is proportional to the sum of variances in the p coordinate directions. Minimizing J, then, results in minimizing the total cluster variance.

Therefore, the problem is to find the optimal (U^*,v^*) which minimizes J(U,v).

<u>Problem:</u> Minimize: $J(U,v) = \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik} f(d_{ik})$

Subject to: $U \epsilon Q_c$.

$v \epsilon R^p$

However, because of the discreteness of Q_c this is not an easy problem to solve. Firstly, because J(U,v) is discontinuous, the continuous function definition of a local minimum does not apply. Consequently, this problem becomes a combinatorial optimization problem. Secondly, because of the large size of Q_c, even for relatively small numbers of elements and clusters, the process of finding the minimum $J(U,v)$ for all pairs (U,v) is exhaustive. Bezdek (1981) concludes that:

> "Although finiteness is sometimes an advantage, it is clear that the size of Q_c will impede search by exhaustion for "optimal" partitionings." (Bezdek, 1981, p. 29)

Algorithms have been developed for approximating the minimum of $J(U,v)$. The most extensively used of these is the ISODATA METHOD. This algorithm assigns each element to a cluster such that the distance between the element and the current cluster centre is smaller than the distance between the element and the current centre of any other cluster. New centres are then calculated, and the process is repeated until there is little change between iterations. There is no known proof of convergence for this iterative procedure (Bezdek, 1980).

Following is the detailed "Isodata Algorithm" as described in Bezdek (1981).

<u>ISODATA</u> <u>ALGORITHM</u> (Duda and Hart)

1. Choose c, $2 \leq c < n$
2. Initialize $U^{(0)} \epsilon Q_c$
3. Initialize the counter $t=0$
4. Calculate the c cluster centres $\{v_i^{(t)}\}$ using equation (1) and $U^{(t)}$
5. Update $U^{(t)}$ to $U^{(t+1)}$ by the following membership criteria:

$$\mu_{ik}^{(t+1)} = 1, \; d_{ik}^{(t)} = \min\{d_{jk}^{(t)}\} \quad 1 \leq j \leq c$$
$$0, \text{ otherwise}$$

This ensures that the object is a member to the cluster whose centre it is nearest.

6. Compare $U^{(t)}$ to $U^{(t+1)}$, if $\|U^{(t+1)}-U^{(t)}\|<\epsilon_L$ stop; otherwise let t=t+1 and go to step 4.

A "tie-breaking" rule is required in the algorithm to account for the occurrence of a point being situated equal distances between two cluster centres. Usually if this occurs the point is assigned to the first cluster in which it was "nearest".

Also, it should be noted that rather than initializing the $U^{(0)}$ matrix, one could initialize the cluster centres $\{v_i^{(0)}\}$ in the algorithm. This would result in the change of clusters centres being used for the termination criteria for the algorithm.

i.e. $\|v_i^{(t+1)}-v_i^{(t)}\|\leq\epsilon_L$

3.3.4 The Fuzzy Clustering Algorithm

Bezdek (1981) again has defined the following criteria which must be met if the fuzzy matrix $U=[\mu_{ik}]\epsilon V_{cn}$ is a solution for the Fuzzy Cluster analysis problem.

1. The grade of membership for an element x_k, into the i^{th} cluster will be greater than or equal to zero, but less than or equal to one.

   $\mu_{ik}\epsilon[0,1] \quad 1\leq i\leq c \quad 1\leq k\leq n$

2. Membership values are chosen so that their sum, for a particular element x_k, is equal to 1 (the sum of all values in a column is 1).

$\sum_{i=1}^{c}\mu_{ik}=1 \quad 1\le k\le n$

3. No cluster is empty, nor contains all elements of X.

   $0< \sum_{k=1}^{n}\mu_{ik}<n \quad 1\le i\le c$

   For computational tractability the third criterion is relaxed by allowing clusters to contain either no elements or all the elements.

   $0\le \sum_{k=1}^{n}\mu_{ik}\le n \quad 1\le i\le c$

The set of all matrices which then define the fuzzy c-partition space for X are defined as:

$$Q_f = \{U\epsilon V_{cn}|\mu_{ik}\epsilon[0,1]\forall i,k;\ \sum_{i=1}^{c}\mu_{ik}=1\forall k;\ 0\le \sum_{k=1}^{n}\mu_{ik}\le n\forall i\}$$

A fuzzy cluster objective function, analagous to that for the crisp case, is defined as follows:

$$J(U,v)=\sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ik})^{s}f(d_{ik})$$

where:

$$U = [\mu_{ik}]\epsilon Q_f$$

$$v = (v_1,....,v_c),\ v_i\epsilon R^P$$

$$f(d_{ik}) = d_{ik}^{\ 2} = (x_k-v_i)^TM(x_k-v_i)$$

$$s\ \epsilon\ [1,\infty\}$$

Raising the grade of membership term, μ_{ik}, to the power "s" is one way in which this objective function differs from the classical WGSS objective function. This exponent "s" is referred to as the exponential weight. The larger its value, the less influence points with low memberships have on determining the cluster centres. The influence of outliers on the determination of the cluster centres is therefore reduced.

Bezdek states that "as $s \rightarrow 1$, fuzzy c-means converges in theory to a "generalized" hard c-means solution" (Bezdek, 1981, p. 70). It approaches a "generalized" hard c-means solution rather than the WGSS solution because the fuzzy objective function includes the option of scaling all distances by the values contained in the matrix M. Bezdek's statement is supported by Dunn (1974) who through experimentation found that as $s \rightarrow 1$ the clustering solution was non-fuzzy even if clusters were completely absent from the data. Conversely, as $s \rightarrow \infty$, the membership assignments approach maximum fuzziness, that is, $[U] \rightarrow [\frac{1}{c}]$.

Because there is an infinite number of choices for s, an infinite family of fuzzy clustering algorithms - one for each s - are defined. At this time no rule for choosing s exists. The value two is often used; however, its choice is arbitrary (Zimmerman, 1985). Bezdek, Ehrlich and Full (1984) suggest that for most data a value between 1.5 and 3.0 will give good results.

As mentioned above, this fuzzy clustering objective function also differs from the classical WGSS objective function in that it allows the user to choose the dissimilarity metric M. For this reason, the fuzzy clustering objective function is a more generalized function.

The clustering problem is then again reduced to finding the optimal pair (U^*, v^*) to minimize the objective function.

Problem: Minimize: $J(U,v)=\sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ik})^{s}f(d_{ik})$

Subject to: $U\epsilon Q_f$

$v\epsilon R^P$

$s>1$

One of the advantages of fuzziness is that because of the continuous nature of J(U,v), it is differentiable with respect to the independent variables μ_{ik} and v_i. A minimum function value, which has the property $\frac{\partial J}{\partial\mu_{ik}\partial v_i}=0$, is therefore exactly defined. Bezdek (1981), by using differential calculus, determined that the optimal pair (U^*,v^*) for the minimum J(U,v) is calculated as follows:

$$v_i^* = \frac{\sum_{k=1}^{n}(\mu_{ik})^s x_k}{\sum_{k=1}^{n}(\mu_{ik})^s} \qquad 1\leq i\leq c \tag{2}$$

$$\mu_{ik}^* = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{s-1}}} \qquad 1\leq i\leq c \quad 1\leq k\leq n \tag{3}$$

An alternative definition for μ_{ik}^* is required for the case when $x_k=v_i$, that is, when $d_{ik}=0$. This case is referred to as "singularity" and when it occurs membership for that particular element x_k in the cluster whose centre is defined by v_i must be equal to 1. Because of the column constraint, membership in any other cluster is then not allowed. To account for this, an additional step in the fuzzy clustering algorithm is included. It is:

$$\text{If } x_k=v_i \qquad \mu_{ik} = 1 \text{ for } i=k \tag{4}$$

0 for $i \neq k$

It should be noted that because of machine round-off this condition rarely occurs.

Although the mathematical derivation for the Fuzzy Cluster Method differs substantially from the crisp cluster method, the resulting fuzzy cluster algorithm is very similar. Again elements are assigned to clusters, this time using a grade of membership. Centroids are calculated for each cluster and new membership values are calculated. This process is repeated until there is little change between iterations. The formal Fuzzy Cluster Algorithm is detailed below (Bezdek, 1981):

FUZZY ISODATA ALGORITHM

1. Choose:
   a. c, $2 \leq c < n$
   b. inner product norm metric matrix, M
   c. s, $1 \leq s < \infty$
2. Initialize $U^{(0)} \epsilon Q_F$
3. Initialize the counter $t=0$
4. Calculate the c cluster centres $\{v_i^{(t)}\}$ using equation (2) and $U^{(t)}$
5. Update $U^{(t)}$ to $U^{(t+1)}$ using equations (3) and (4) and $\{v_i^{(t)}\}$
6. Compare $U^{(t)}$ to $U^{(t+1)}$, if $\|U^{(t+1)}-U^{(t)}\| \leq \epsilon_L$ stop; otherwise let $t=t+1$ and go to step 4

At this point it is worth looking at the arbitrarily chosen M matrix more carefully. M influences the shape of the cluster which is determined by the algorithm. For the most common case, where M=[I], the clusters identified tend to be spherical (Windham, 1983). Other frequently used norms are (Dunn, 1974):

1. $M=[\mathrm{diag}(\sigma_j^{2})]^{-1}$, σ_j^{2}=sample variance of the j^{th} dimension of vectors $x \epsilon X$; for dimensions 1 to p.

   Here the "distance" in each dimension is scaled by a factor that reflects the spread of values for that dimension. This results in clusters which are compact in relative terms. For example, if two dimensions P_1 and P_2 exist, and their points vary as follows;

   dimension P_1 [. . . . .]
   dimension P_2 [.......]

   it is clear that two points which are considered close together in P_1 would be considered relatively far apart in P_2. Scaling the distances by the inverse of the variance normalizes the distances in each dimension before comparing them.

2. $M=[COV(x)]^{-1}$=sample covariance matrix for values in each dimension p for all x_k.

   This scaling matrix accounts for variance differences in the same way the above scaling matrix does, and also attempts to decrease the effects of statistical dependence between variables (Bezdek, 1981).

Windham (1983) introduces an algorithm which, as well as optimizing U and v, optimizes the metric M within each cluster. That is, he attempts to choose optimal U^*,v^* and M^* which minimize

$$\sum_{i=1}^{c}(\mu_{ik})^{s}(x_k-v_i)^{T}M_i(x_k-v_i)$$

This algorithm identifies ellipsoidally shaped clusters.

Finally, one must consider the convergence properties of this FUZZY ISODATA algorithm. The iteration method used in this algorithm does not guarantee convergence to a globally optimum solution. At best, Bezdek (1980) concludes that this iterative procedure "always terminates at a local minimum, or at worst, always contains a subsequence which converges to a local minimum of the generalized least squares objective function which defines the problem" (Bezdek, 1980, p. 1). It is therefore desirable to perform the iterative procedure from several different initial membership matrices. If the same solution is obtained from each of these different starting positions one may then assume, with reasonable confidence, that an optimum solution has been found.

Dunn (1973) performed many numerical experiments and concluded that the Fuzzy Cluster Method converged rapidly to an optimal partition from virtually all starting guesses, $U^{(0)}$, when relatively crisp and well-separated clusters were present in the data. When crisp and well-separated clusters were not present in the data, the procedure still converged to a partition which was truly fuzzy (i.e. the resulting membership functions departed significantly from the hard limits 0); however, the convergence was slower.

3.4 Cluster Validity

Having determined the "optimum" cluster structure by the method just described, a measure of how good the solution is, is required. The clustering algorithm itself establishes the optimum values for the cluster centres, v_i's, and the membership matrix, [U], for a given number of clusters, c, and exponential weight, s. It is the question, then, of what are the best values for c and s that must be addressed.

Initially the value of the objective function itself was suggested as a measure for comparing results obtained using different exponential weights or number of clusters. It was proposed that the objective function be calculated for each solution, with the results associated with the overall minimum function considered as "optimum". Bezdek (1981) disagrees with this approach suggesting that the overall minimum J does not necessarily yield the "best" clusters when c and s vary.

A number of scalar measures were therefore developed to reflect the quality of specific cluster structures. None of these is without limitations, nor does one claim to be more accurate than the others. The two measures that were used in this esearch were the partition coefficient and the entropy measure. Both of these measures represent the degree of fuzziness of a solution. The premise for measuring cluster validity is that better solutions are less fuzzy.

Bezdek (1981) provides the following definition of the partition coefficient. If $U \epsilon Q_f$ is the fuzzy c-partition for n data points, the partition coefficient of U is defined as the scalar

$$F(U;c) = \sum_{k=1}^{n} \sum_{i=1}^{c} \frac{(\mu_{ik})^2}{n}$$

Because of the form of this function a global minimum value for F is guaranteed. This minimum value occurs at the state of maximum fuzziness for the system, i.e. when $[U]=[\frac{1}{c}]$. The function attains a maximum value of 1.0 for all hard c-partitions of X. Bezdek (1981) summarizes the properties of the partition coefficient, F(U;c), as follows:

1. $\frac{1}{c} \leq F(U;c) \leq 1$
2. $F(U;c)=1 \leftrightarrow U \epsilon Q_f$ is hard
3. $F(U;c)=\frac{1}{c} \leftrightarrow U=[\frac{1}{c}]$

Maximizing F(U;c) minimizes the amount of fuzziness in a solution. A formal strategy for determining the most valid clustering structure for different values of c and s is then as follows:

1. Using the fuzzy c-means clustering algorithm, determine at each c=2,3,....,n-1 one or more optimal c-partitions of X (in general, also varying s).
2. Let Ω_c denote the finite set of optimal candidates, for each c, as determined in 1.
3. solve by direct search

   $\max_c\{\max_{\Omega_c}\{F(U;c)\}\}$

The solution $(U^*,c^*)$is considered to be the most valid clustering of X.

The other cluster validity measure is based on the concept of entropy. Entropy is a measure used to describe the statistical uncertainty associated with a given state of a system. The concept was first introduced by Shannon in 1948. According to Bezdek (1981), Shannon reasoned that the unique vector $(\frac{1}{c})$ represented the state of maximum uncertainty and thus suggested that the entropy measure should maximize at $(\frac{1}{c})$. Conversely if the system state was crisp, there was no uncertainty and consequently the entropy measure should be a minimum. Because fuzzy sets represent states of uncertainty, the entropy measure was a natural choice for measuring degree of fuzziness.

De Luca and Termini (1972) first introduced the entropy measure for the fuzzy 2-partition case. Bezdek (1981) defines a generalized entropy measure for any fuzzy c-partition as:

$$H(U;c) = -\frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik} \ln(\mu_{ik})$$

where $\mu_{ik}\ln(\mu_{ik})=0$ whenever $\mu_{ik}=0$. This measure ranges from zero to ln(c). A value of zero occurs when the partitioning has minimum fuzziness, when the partitions are crisp; the maximum value, ln(c), occurs when the partitioning of X is "most fuzzy", when $U=[\frac{1}{c}]$. These properties are summarized below;

1. $0 \le H(U;c) \le \ln(c)$
2. $H(U;c)=0 \leftrightarrow U \epsilon Q_f$ is hard
3. $H(U;c)=\ln(c) \leftrightarrow U=[\frac{1}{c}]$

Minimizing H(U;c) minimizes the amount of fuzziness in a solution, and therefore suggests another measure for cluster validity. A formal stategy analagous to that for the partition coefficient may be adopted for determining the most valid clustering structure. Specifically,

1. Using the fuzzy c-means clustering algorithm, determine at each c=2,3,....,n-1, one or more optimal c-partitions of X (in general, also varying s).
2. Let Ω_c denote the finite set of optimal candidates, for each c, as determined 1.
3. Solve be direct search

   $\min_c\{\min_{\Omega_c}\{H(U;c)\}\}$

The solution, (U^*,c^*) is considered to be the most valid clustering of X.

Both the partition coefficient and entropy measures are limited by (1) their monoticity and (2) their lack of declaration of a suitable benchmark for determining when a cluster hypothesis should be totally rejected. The monoticity property tends to indicate that the most valid partition is the 2-partition. If both measures indicate that the 2 cluster solution is the best solution the trends of the measure values should be examined. Zimmerman (1985) suggests, for the partition entropy measure, that one choose "the i^*-partition for which the value H(U;c) lies below the trend when going from c^*-1 to c^*" as the best solution. Conversely, one could study the trend of the partition coefficients and choose the partition for which F(U;c) lies above the trend when going from c^*-1 to c^*.

3.5 The Fuzzy C-Means (FCM) Program

3.5.1 The Program Structure

The FCM program is a FORTRAN program written by William E. Full of Wichita State University, Wichita, Kansas. Bezdek, Ehrlich and Full (1984) provide the FORTRAN source code in the paper "FCM: The Fuzzy C-Means Clustering Algorithm". The program is based on Bezdek's Fuzzy C-Means algorithm which is described in detail in Section 3.3.4. Only the portion of the program which does not directly parallel this algorithm is discussed in this section.

The FCM program structure is illustrated in the flowchart in Figure 3.1. A listing of the FORTRAN coding of the program has been included in Appendix A.

3.5.2 Input Variables

Following are the variables that must be input into the FCM program. The required input format is described in the FCM program listing.

The first variable is the ICON variable. This variable identifies the scaling matrix, M. The user is presented with three choices for this matrix:

ICON=1 → $M=[I]$

ICON=2 → $M=[\mathrm{diag}(\sigma_j^2)]^{-1}$

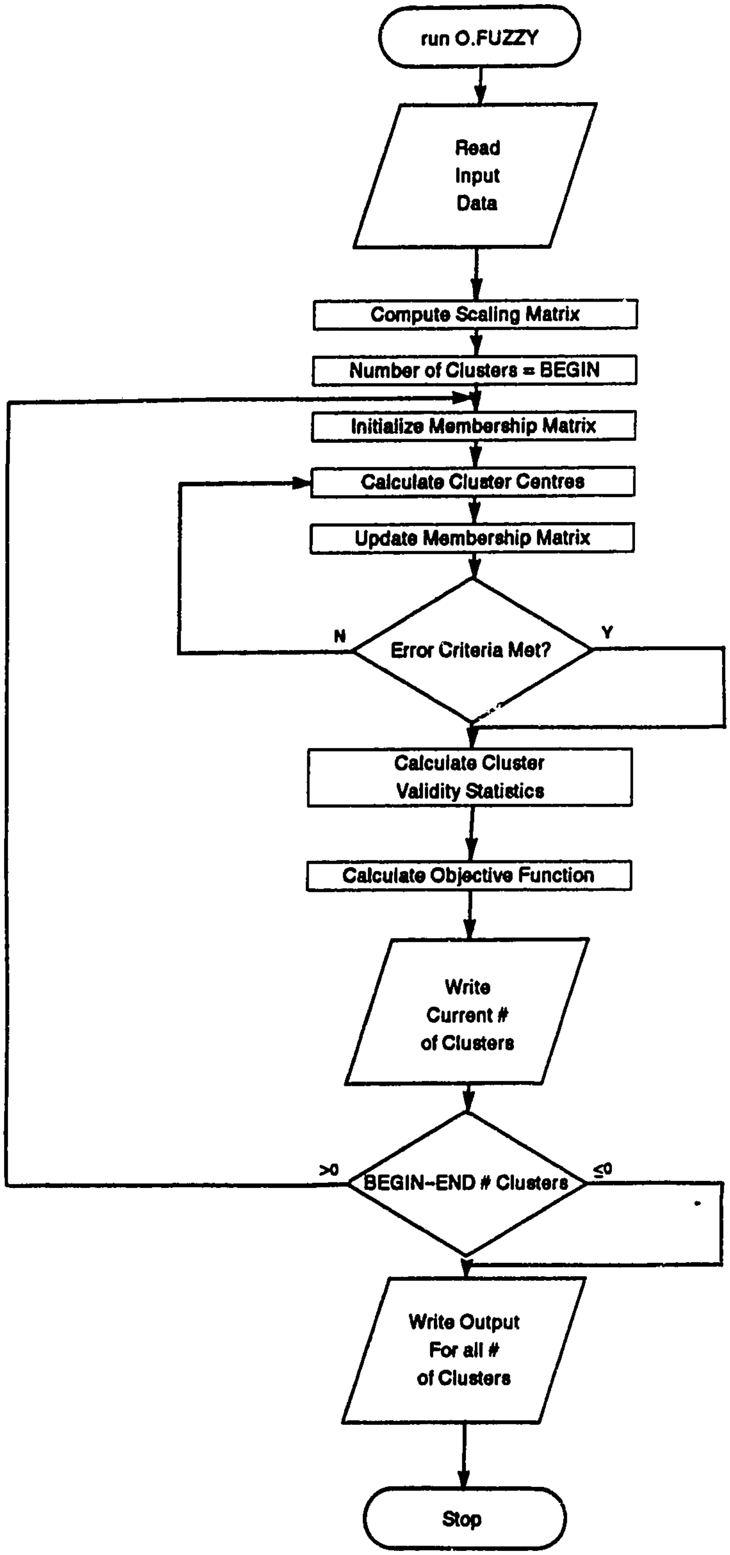

run O.FUZZY
Read Input Data
Compute Scaling Matrix
Number of Clusters = BEGIN
Initialize Membership Matrix
Calculate Cluster Centres
Update Membership Matrix
N
Error Criteria Met?
Y
Calculate Cluster Validity Statistics
Calculate Objective Function
Write Current # of Clusters
>0
BEGIN-END # Clusters
≤0
Write Output For all # of Clusters
Stop

Figure 3.1 Fuzzy C-Means (FCM) Clustering Algorithm Flowchart

ICON=3 → $M=[COV(x)]^{-1}$

For more detailed information regarding these matrices see Section 3.3.4.

The next parameters specified by the user are:

QQ=weighting exponent "s" ($1<s<\infty$)

KBEGIN=starting number of clusters for the program run

KCEASE=finishing number of clusters for the program run

(KCEASE≥KBEGIN)

Note that the program allows the user to input a range of "number of clusters" for which an optimal solution, (U^*,v^*), is calculated for each.

Lastly, the object data are input.

3.5.3 Initial Cluster Membership Values

As was discussed in Section 3.3.4, the Fuzzy Cluster algorithm does not guarantee convergence to a globally optimum solution. It is therefore desirable to run the program a number of times beginning at several different initial cluster membership matrix values ($U^{(0)}$'s). The FCM program provides two different methods for initializing the cluster membership matrix. Between these two methods an infinite number of initial membership matrices can be created. The choice of which method is to be used must be specified internally in the program. To understand how this option is activated the membership initialization portion of the program must first be examined.

The assignment of initial cluster membership values is subject to the constraint that the sum of all membership values, for an object into all clusters, must be equal to one. For each object the program assigns initial membership values using a looping process which for the first cluster membership takes a proportion of the value 1.0, and then continues taking a proportion of the remaining amount to be proportioned until the number of clusters less one (NCLUS-1) are apportioned. The membership value for the final cluster is the remaining amount. This process ensures the sum of all memberships for an object is one. All objects are apportioned membership values in this manner.

The proportioning process is done by multiplying the value 1.0 in the first instance, and the remaining portion in all other instances by a number which is greater than zero and less than one. This number is a function of the total number of clusters for the program run, the current cluster number whose membership is being apportioned, and a variable called RANDOM. The last two variables change for each cluster membership calculation for a particular object.

Two methods exist for specifying the value of the RANDOM variable, allowing the user to vary the initialization method. One method simply assigns the value of the RANDOM variable at the beginning of the initialization process equal to 0.7731. For each subsequent membership calculation, for all objects, this value is then halved. The other method assigns a pseudo-random number for the RANDOM variable for each membership calculation. The random variable is assigned using the random number generator subroutine RANDU from the IBM Scientific Subroutine Package (SSP). This subroutine computes pseudo random numbers between zero and one. Again, for each membership calculation this value is then halved. It should be noted that with this method an infinite number of initial membership matrices could be generated simply by changing the value of the random number generator seed number in the FCM program listing. The seed number must be an odd integer with nine or less digits (SSP).

The FCM program, in its original state, will initialize the membership matrix using the first method described above. The program statement which would call the RANDU subroutine is not activated because it is designated as a comment statement in the program listing. The comment designation occurs because of the presence of a "C" in the first column of the FORTRAN statement. Therefore, if the user desires initializaton by the first method he makes no adjustments to the program. If he desires initialization by the second method he must eliminate the comment designation in the RANDU call statement before program compilation.

3.5.4 Termination Criteria

Termination of the Fuzzy C-Means algorithm occurs when the maximum change in cluster membership values, for all objects, between iterations, is less than or equal to some predefined, internally specified, termination criterion (ϵ_L). In the FCM program this value has been set equal to 0.01. As well, a variable LMAX is defined in the program which limits the number of iterations allowed without the termination criterion being met. In the program LMAX is equal to 50 iterations.

For this research the termination values as specified in the FCM program wer [illegible] d. Bezdek, Ehrlich, and Full (1984) state that lower [illegible] termination criteria almost always results in more iterations to program termination. The values defined in the program are deemed as a good balance between program running costs and desired result accuracy.

3.5.5 Cluster Validity Statistics

Upon termination of the Fuzzy C-Means algorithm, for a specified number of clusters, the cluster validity statistics and the objective function are computed. The cluster validity statistics calculated in this program are the partition coefficient, $F(U;c)$, and the entropy measure, $H(U;c)$. These statistics are discussed in detail in Section 3.4. The number of iterations required for the Fuzzy C-Means algorithm's convergence is also monitored.

The values of these statistics are printed, for the current number of clusters, at the termination of the Fuzzy C-Means algorithm. A summary of these statistics, for the range of cluster values specified in the input data, is the final output from the FCM program.

3.5.6 Miscellaneous

It is worth noting that the FCM program does not account for the possibility of an object and a cluster centre having the exact same location, the "singularity" case described in Section 3.3.4. Bezdek, Ehrlich, and Full (1984) acknowledge this omission and state that to their knowledge this event has never occured in nearly ten years of computing experience. Thus, for practical purposes this problem was disregarded.

3.5.7 Program Changes

A few minor changes to the program's output format were made. Changes were made to improve the appearance and to decrease the size of the output files. The size of the output files were decreased by eliminating the program step which reiterated the object input data, specifically, the object dimension values.

The original program is stored in the file S.FUZZY.1 and the program which contains these described changes is stored in the file S.FUZZY.2. Both files are stored on magnetic tape, filed in the University of Alberta magnetic tape library as volume 127505, rack number 071366.

4. THE DATABASE

This chapter describes the variables that were used in the cluster analysis procedure. For ease of description these variables have been divided into the following variable "types": perceived variables, objective variables, and socioeconomic variables.

The perceived variables included both the travelers' perceived walk times from their homes to their transit stops and their perceived wait times at these stops. The corresponding actual travel time values and a composite utility value for the transit service available to each individual in the sample were included as objective variables. Although the composite utility value measures the overall satisfaction the user associates with the transit network, a somewhat perceptual measure, it is based on objective measures and actual observed behaviour and therefore was classified as an objective variable. This measure will be discussed in more detail in Section 4.3.4. Socioeconomic variables included gender, age, income, job type and usual mode choice.

4.1 The Data

The cluster analysis data were derived from two sources: (1) The 1983 Morning Commuter Survey (contained in the CMCS data file) and (2) a file containing transit route alternative information for the calibration of logit mode split and public transit route choice models (D.BS.MDM1).

The CMCS file contains actual survey responses, and data derived from these responses. A full description of the CMCS database is contained in Working Paper 7 by Hunt (1984).

A number of the 1702 complete CMCS interviews were investigated in further detail by Hunt (1988) and Cooper (1989). Some of this additional information was used to develop logit mode split and public transport route choice models. For the calibration of these models, choice sets describing all relevant transit options for an individual's home-to-work trip were required. For each transit alternative the following information was compiled:

- The walk distance from the individual's home to the public transit stop
- The total in-vehicle travel time
- The total number of transfers and the total transfer waiting time
- The frequency of transit service
- The walk distance from the destination end public transit stop to the individual's place of employment.

In addition to transit route alternative information these choice sets also included the individual's age, sex, and income. The total number of interviews for which these choice sets were developed was 638. All choice set information is stored in the computer file D.BS.MDM1.

Because transit information already existed for these 638 interviews, they became the base for the cluster analysis database. This database size was deemed acceptable, for this research, for the following reasons:

1. The sample is regarded as a large sample.
2. The generation of all public transit alternatives, and the information required for these alternatives, for more individuals would be overly demanding.

It should be noted that not all of these 638 interviews were complete. That is, not all individuals answered the interview questions whose answers were part of the database for this study. Generally when information was not attained zero values were indicated in the data files. It was felt that these zero values could have a significant effect on the final cluster results and therefore incomplete interviews were disregarded. This will be discussed in more detail later in this chapter.

4.1.1 Data Reliability

As was stated in Chapter 1, over 1700 individuals who worked in the Edmonton CBD were interviewed regarding various aspects of their socioeconomic character and home-to-work trip. This information was stored in the Coded Morning Commuter Survey (CMCS) data file. With regard to the reliability of this data, Hunt (1988) states,

> "Because of the detailed and specific nature of the set of instructions for the interviews, and the capabilities of the interviewers, it is judged that the information obtained in the survey is highly

credible." (Hunt, 1988, p. 4-19)

Since the data for this research were largely derived from the CMCS data, they too are judged as highly credible.

4.1.2 Data Management

The University of Alberta computing facilities were used for all data management functions and data storage. All files have been stored on magnetic tape filed in the University of Alberta magnetic tape library as volume 127505, rack number 071366. All data manipulation programs written for this research are indexed in Appendix H. A flowchart of the database development is also included in this appendix.

4.2 Perceived Variables

4.2.1 Perceived Walk and Wait Times

For this study, survey respondents' reported values of public transit walk and wait times were used to represent their perceptions of these attributes (see Section 2.1).

In the 1983 Morning Commuter Survey each individual who regularily chose public transit for his morning home-to-work trip was asked:

1. How long it took to walk from his home to the bus stop (LRT station)
2. How long he usually waited for the bus (LRT).

Those persons who did not choose public transit were asked to suppose that they were going to choose the bus for their home-to-work trip. Each was then asked:

1. How long he thought it would take to walk to the bus stop
2. How long he thought he would wait for the bus.

All reported data are stored in the CMCS computer file.

4.3 Objective Variables

4.3.1 Actual Walk and Wait Times

In order to compare perceived walk and wait times between individuals, corresponding actual walk and wait time values were desired. In this study, however, true values of walk and wait times were unavailable. For the walk time variable what was available was the walk distance from the individual's home to his origin-end public transport stop (all walking distances are shortest path distances measured from city maps). Using the measured walk distance and an assumed average walking speed a walk time value for every individual could be estimated. However for the cluster procedure, it is not the absolute values of the variables that are important but rather the relative distance they are apart from one another. Multiplying all the walking distances by the same number does not change the relative distances between objects in this dimension. It is, therefore, simply more efficient to use the walk distances themselves in the clustering procedure.

Wait time models generally estimate waiting time as a function of the public transport service frequency (O'Farrell and Markham, 1974; Algers et al, 1975; Chapman, Gault and Jenkins; 1976; Bovy and Jansen, 1979). For routes where vehicle arrivals are frequent, the models are based on the assumptions that passengers arrive at random and that they board the first vehicle that arrives. When headways between vehicles are longer, the models are based on the assumption that passengers will systematically arrive at the stop so as to avoid long waiting times (Algers et al, 1975). Therefore different models may exist depending on the transit vehicle frequency. A typical wait time model is defined by Bovy and Jansen (1979). Their model, which is based on empirical investigations consists of three separate functions. They are:

1. $\bar{w} = \bar{h}$ for very high frequency service ($\bar{h}<2$ min)
2. $\bar{w} = \frac{1}{2}\bar{h}$ for medium frequency service ($2\leq\bar{h}\leq10$ to 15 min)
3. $\bar{w} = \alpha\bar{h}$ where α is small, e.g. 0.1 to 0.2, for low frequency service ($\bar{h}>15$ min)

where:

$\bar{w}$ is the average wait time

$\bar{h}$ is the average headway

These models assume perfect schedule adherence by the public transport vehicles.

Generally, for the 1983 Edmonton situation both public transport alternatives with low and medium frequencies (models 2 and 3) existed (public transport frequencies were determined from Edmonton Transit schedules). There is, to the author's knowledge, no known calibration of an Edmonton model for low frequency routes; consequently, the wait time for all passengers was assumed to be equal to one half of the headway. This assumption likely resulted in actual wait time values that were too high for public transport alternatives with larger headways. This will be discussed in further detail in Section 6.3.3.

If one assumes that the actual wait time is equal to one half of the vehicle headway then the same argument as was used for the walk time variable can be applied. That is, since the wait times are a linear function of the public transport service frequency, using the frequency values as proxies for the wait times will yield the same results as using the estimated wait times in the cluster analysis. Therefore, for the cluster analysis procedure public transport service frequencies were used to represent the actual wait times.

Having determined that the walk distance would represent the walk time variable and the service frequency would represent the wait time variable, the values of these variables had to be determined for all individuals in the research database. Both the origin-end walk distances and the public transport service frequencies were stored as part of the public transport attribute information, for all public transport alternatives available to an individual, in

D.BS.MDM1.

For the transit users the procedure to determine the objective variable values was then as follows. First, the public transit alternative that was actually chosen by the individual was identified. Then, the origin-end walk distance and service frequency for this chosen alternative were extracted from D.BS.MDM1.

For car users the process was not so straightforward. Before the public transit walk distances and service frequencies could be extracted from D.BS.MDM1, the public transit alternative that each individual would have most likely chosen had to be identified. A multinomial logit route choice model, the TRAM model developed in Edmonton by Cooper (1989), was employed for this task. A brief description of this model and the theory on which it is based is given below.

4.3.2 Multinomial Logit Theory

The following section briefly introduces multinomial logit theory. The discussion, unless otherwise specified, is primarily derived from Domencich and McFadden (1975).

The basis of the logit model is that individuals, rather than aggregate zonal populations, represent the basic decision unit. The disaggregate model is based on the Theory of Rational Choice Behaviour which states that given a number of alternatives the individual will choose the alternative that he finds most desirable. This "desirability" will depend on the attributes of the alternative and the socioeconomic characteristics of the individual. In mathematical terms, each individual, i, associates with each alternative, j, a utility function, U, which measures the desirability of an alternative, x, for the specific socioeconomic characteristics of the individual, s.

$$U(j,i) = F(x_j,s_i)$$

The individual will then choose the alternative that maximizes his utility.

Because it is not completely understood how individuals behave, each utility function also includes an error term. Therefore, the utility for a particular alternative, for an individual drawn randomly from the population is written as:

$$U(j,i) = V(x,s) + E(j,i)$$

where V(x,s) is the representative utility common to all members of the population, and E(j,i) is the random component. Thus, the probability of choosing alternative j^*, for individual i is,

$P(j^*,i)=P(U(j^*,i)>U(j,i))$ for all alternatives J, except j^*

The assumption that E(j,i) is distributed according to the Weibull distribution results in the Logit Choice Probability Model. This model has the form:

$$P(j^*,i) = \frac{e^{V(j^*,i)}}{\sum_{j=1}^{J} e^{V(j,i)}}$$

The method of disaggregate modeling, then, is to specify the utility function for each alternative. This function is a linear function of the form $\Sigma\, \phi_n x_n$, where x_n are the attributes for each alternative, and ϕ_n, are the vector of coefficients. It is these coefficients which must be calibrated to the data.

Calibration of the logit model employs the Maximum Likelihood Technique. This is a statistical method which determines the values of the coefficients such that the model best replicates observations of actual behaviour. Consequently, observations of actual decisions individuals made when in the choice situation are required. All reasonable alternatives available to the individual must be defined. The probability of an individual choosing his actual chosen alternative can then be calculated using the model and some assumed coefficients. The probability of all individuals choosing their actual chosen alternatives is defined as the likelihood. It is calculated as the product of all these probabilities. This is the term the model maximizes.

Basically the maximum likelihood technique adjusts the coefficient values until those which best replicate observations of actual behaviour are determined. For computational ease the natural logarithm of the likelihood term is maximized.

4.3.3 The TRAM Model

The TRAM model is a logit model developed by Cooper (1989) to represent transit route choice behaviour of morning downtown commuters in Edmonton.

For the calibration of this model, choice sets were developed for 121 of the transit users identified from the CMCS data file. As was described earlier, each choice set included both attributes of the transit route alternatives and socioeconomic data describing the individual. The attributes of the transit route alternatives included in the choice sets were:

- The distance from the individual's home location to his "getting-on" public transit stop in metres (DISTO)
- The frequency of public transit service in minutes (FREQT)
- The total amount of time spent waiting for public transit at transfer locations in minutes (TRTIM)
- The number of transfers required (NBTR)
- The total amount of time spent in the public transit vehicle in minutes (TVTIM)

- The distance from the "getting-off" public transit stop to the individual's place of employment in metres (DISTD)

The individual's gender, his age and income categories were also contained in each choice set.

Using the Maximum Likelihood technique, Cooper (1989) found the following utility function best replicated the choice behaviour observed:

$$U_{ji}=-0.00609(DISTO)-0.162(TTOTAL)-0.115(FREQT)-1.84(NBTR)-0.00245(DISTD)$$

where all the variables are defined as above, except:

TTOTAL = total travel time in minutes (riding time + transferring time)

For this research the TRAM model was used to determine which public transit alternatives car drivers would ideally have chosen, so that actual public transit attribute values could be estimated. This required that all public transit alternatives, for each car driver in the database, be assigned utility values using the TRAM model. According to choice behaviour theory, the alternative with the highest utility value is then the alternative that the model predicts will be chosen by the individual. The corresponding origin-end walk distance and service frequency for this alternative could then be extracted from the D.BS.MDM1 file. This procedure was employed for all car users in D.BS.MDM1.

4.3.4 The Composite Utility Variable

Another output from the Logit formulation is the Composite Utility (CU) value. Cooper (1989) used this value as a comparative measure for the analysis of transit network alternatives. It characterizes the overall utility or satisfaction an individual associates with the network of alternatives available to him (Ben-Akiva and Lerman, 1979). With respect to public transit, this concept is illustrated in the following example. Suppose an individual, person 1, lives in a neighbourhood serviced by three transit routes, each which has a service frequency of 15 minutes. Suppose another individual, person 2, lives in another neighbourhood which is serviced by two routes, one with a 30 minute frequency and the other with a 15 minute frequency. From a network point of view, assuming an equal walk distance to all routes, the overall transit service available to person 1 is better than that available to person 2. Person 1, therefore, should be more satisfied with the available transit service. This satisfaction will be reflected by a larger composite utility value.

One may hypothesize, then, that those persons with more satisfaction, or better transit service, may have better attitudes towards public transit. This attitude may be reflected in their perceptions regarding the public transit walk and wait times. For this reason, the composite utility value was included as a variable in the clustering procedure.

Mathematically, the composite utility value is based on the utilities for all the alternatives available to an individual. That is, for each individual (Ben-Akiva and Lerman, 1979):

$$CU=\ln[\sum_{j\epsilon J} e^{U_j}]$$

where:

CU= the composite utility based on the utility values for all alternatives, J, available to the individual

U_i= the utility value for alternative j, $j\epsilon J$, for the individual

4.4 Socioeconomic Variables

4.4.1 Income, Age, and Gender

The format of the 1983 Morning Commuter Survey was such that interviewers wrote the respondents' answers to the questions on an interview form. However, for the income and age questions the individual was asked to indicate, for himself, his age group and income group from a given scale of different groups for each variable. This was to encourage honest reponses (Hunt, 1984). The categories defined by Hunt (1984) were also used for this research. They are:

1. For the income variable, the before tax salary groups, in dollars per year are:

00 - no salary group indicated
01 - salary group question answered with "don't know" by respondent
02 - salary group guessed at by interviewer (guess not recorded)
03 - 0000 - 4,999 dollars per year
04 - 5,000 - 9,999 dollars per year
05 - 10,000 - 14,999 dollars per year
06 - 15,000 - 19,999 dollars per year
07 - 20,000 - 24,999 dollars per year
08 - 25,000 - 29,999 dollars per year
09 - 30,000 - 34,999 dollars per year
10 - 35,000 - 39,999 dollars per year
11 - 40,000 - 44,999 dollars per year
12 - 45,000 - 49,999 dollars per year
13 - 50,000 - 59,999 dollars per year
14 - 60,000 - 69,999 dollars per year
15 - 70,000 and over
20 - student and no salary group indicated
30 - summer employee indicated and no salary group indicated
33 - summer employee and salary group 0000 - 4,999 indicated
50 - part time employee and no salary group indicated

2. For the age variable, the age groups are defined as:

0000 - no age group indicated
1519 - 15 to 19 years
2024 - 20 to 24 years
2529 - 25 to 29 years
3034 - 30 to 34 years
3539 - 35 to 39 years
4044 - 40 to 44 years
4549 - 45 to 49 years
5054 - 50 to 54 years
5559 - 55 to 59 years
6064 - 60 to 64 years
6569 - 65 to 69 years

3. For the gender variable, the categories are defined as:

1 - male
2 - female

As was discussed earlier, interviews which contained unanswered questions were not considered in the clustering procedure. As well, atypical groups, such as those defined for the income variables to account for summer employees and part time employees were not considered. Again, this was because their presence would affect the values of the cluster centres and, therefore, the final results. For example, if a cluster contained mostly individuals from income groups 03, 04, and 50 (generally low income groups), the cluster centre value for the income dimension would not truly reflect the average income in the cluster. The presence of individuals whose income was in the 50 category would artificially inflate the calculated average income value. Therefore, all interviews in the data base which belonged to the income groups 00,01,02,20,30,33, and 50, and to the age group 0000 were not included in the clustering procedure.

The distributions of the gender, income and age variables, for the research database are shown in Figures 4.1, 4.2, and 4.3, respectively. It should be noted that there is probably a larger proportion of females, of lower income individuals, and of twenty to twenty-nine year old individuals in this data sample than there are in the population of individuals making the home-to-work trip to the Edmonton CBD. This is because this data sample has been derived from the CMCS data which also exhibit this trend. Hunt (1984) suggests that this trend is a result of the way in which individuals were selected for interviewing. He states

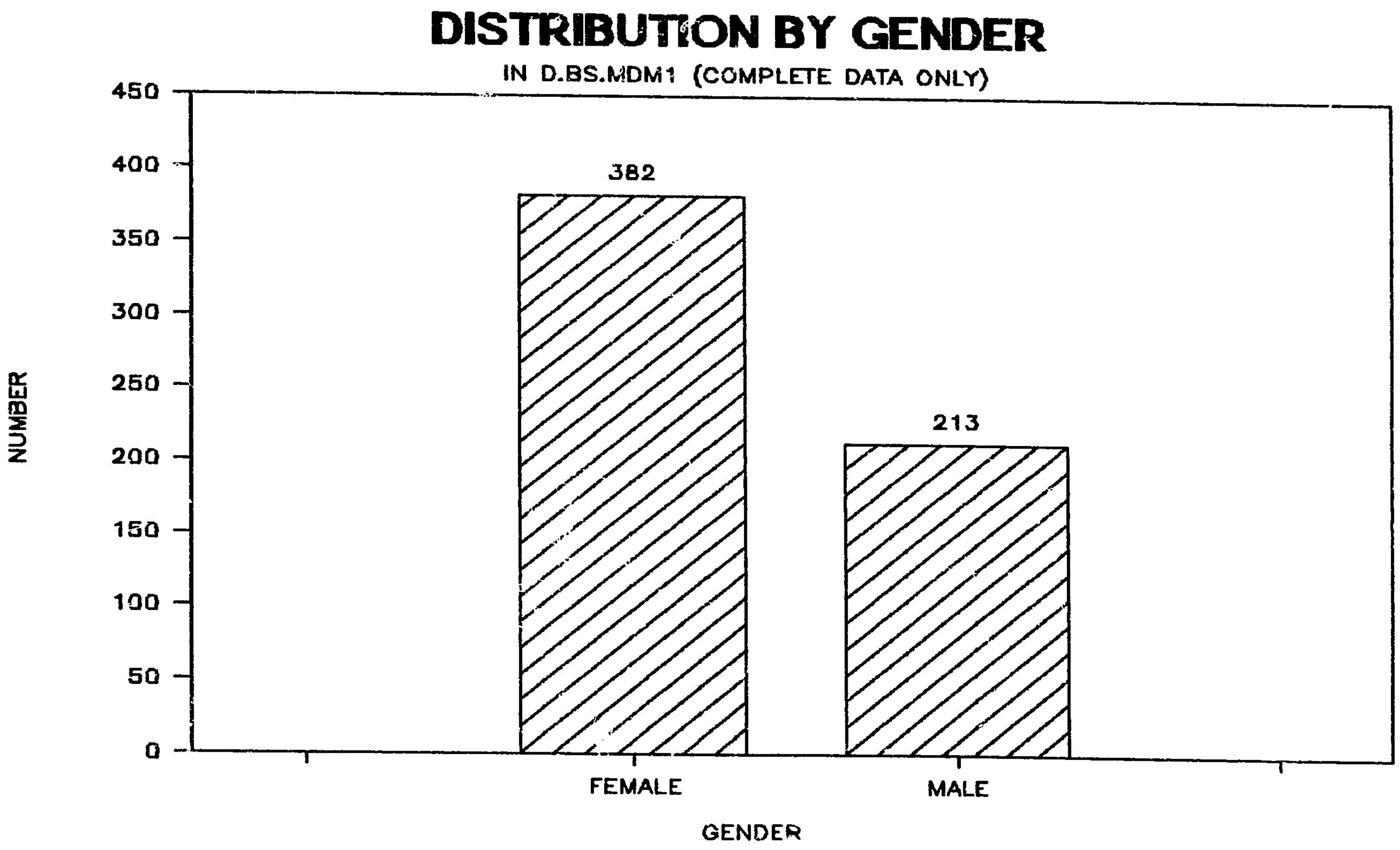


Figure 4.1 Gender Distribution in D.BS.MDM1 (Complete Data Only)

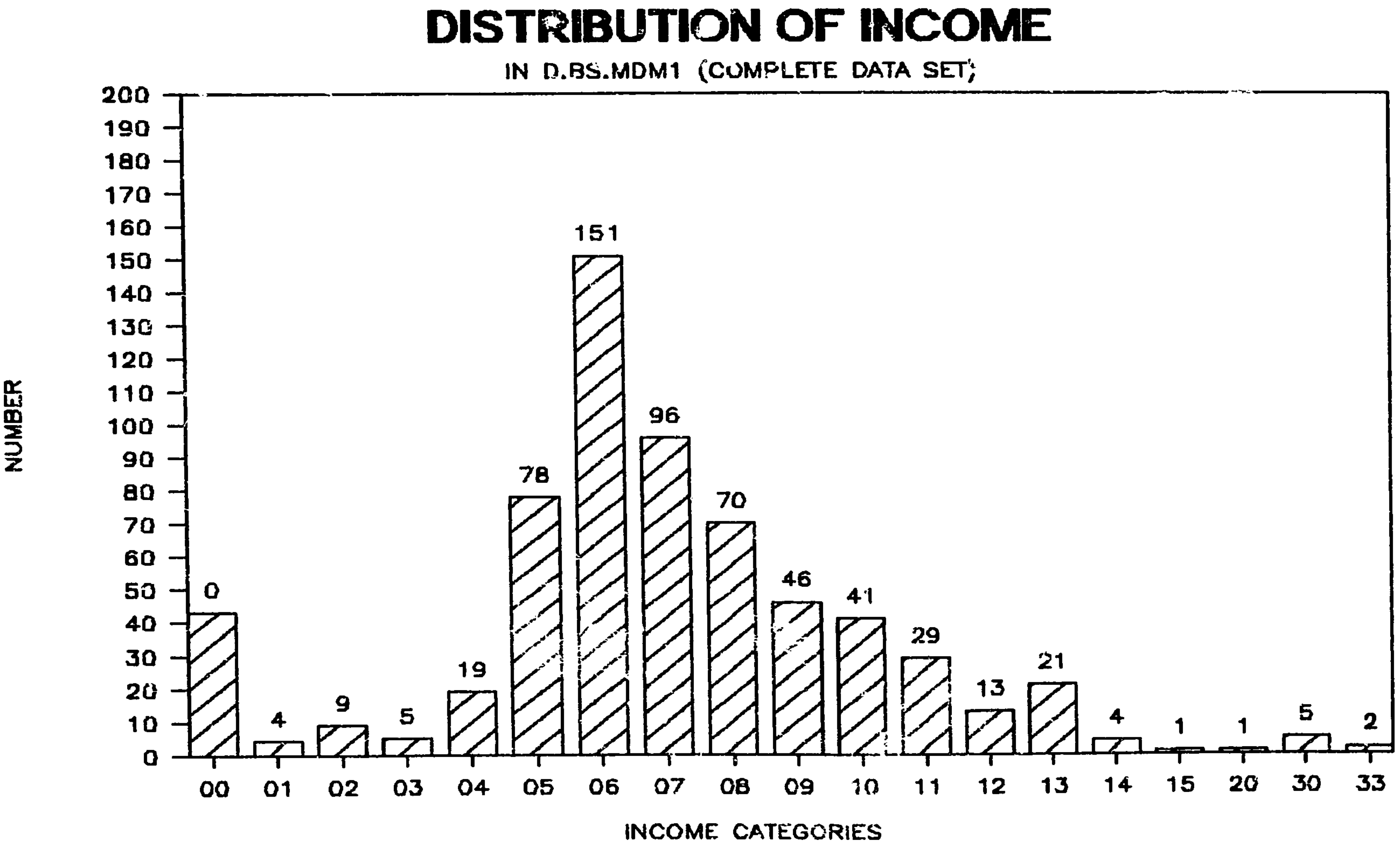


Figure 4.2 Income Distribution in D.BS.MDM1 (Complete Data Only)

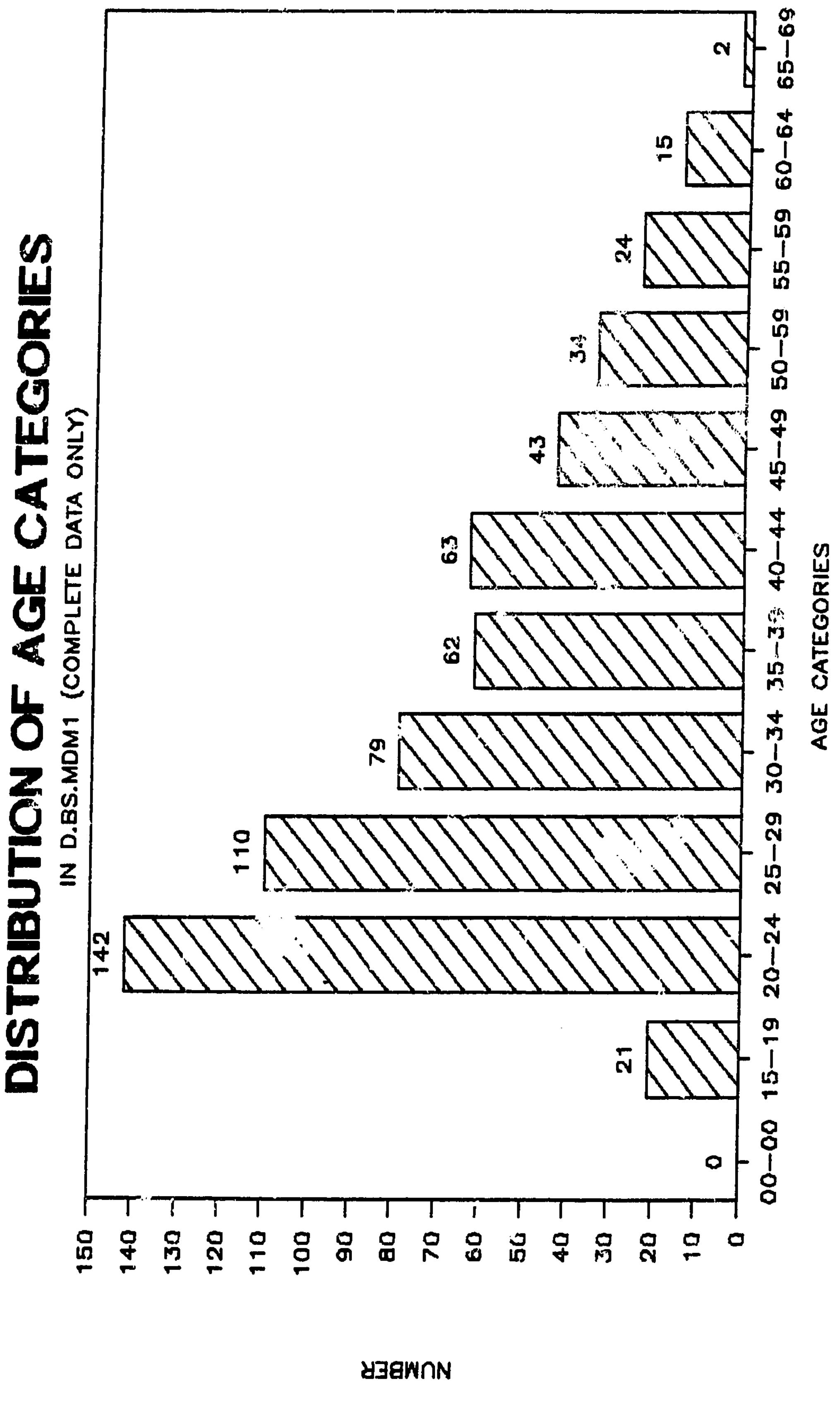


Figure 4.3 Age Distribution in D.BS.MDM1 (Complete Data Only)

that because employers were able to specify which employees were to be interviewed, they were probably less likely to disrupt important, higher paid employees; therefore, the CMCS sample probably includes a larger proportion of lower income persons than exists in the population of persons making the home-to-work trip to the Edmonton CBD.

4.4.2 Job Type

It is felt that in addition to an individual's income the type of job he does may affect how he perceives public transit, more specifically, how he perceives public transit walk and wait times. One would expect the "job type" variable and income variable to be highly correlated.

In the morning commuter survey, each individual was asked to describe his job, his job activities, and to state his job title. This information was used by Hunt (1984) to compile various "job type" descriptions. For example, very precise job descriptions based on the Canadian Classification and Dictionary of Occupations, as well as more general descriptions based on the individuals position within the hierarchy of his place of employment were compiled. For this application it was felt that the categories should be ordered in some fashion, for instance, according to job prestige. Therefore, the individual's position within the hierarchy of his place of employment was chosen as the "job type" variable. The categories, as defined by Hunt (1984), and the membership distribution to each category are given below:

CATEGORY		NUMBER IN CATEGORY
0 -	No indication given	0
1 -	"Owner","Partner","President" "Government Minister", or "Commissioner"	1
2 -	"Vice-President","General Manager" "Deputy-Minister", or "Assistant Deputy-Minister"	5
3 -	"Manager"	51
4 -	"Supervisor" or "Foreman"	53
5 -	All general staff	528

As was the case for the distributions of the other socioeconomic variables, the distribution of job types also includes a large proportion of individuals with "less prestigious" job types.

4.4.3 Usual Mode Choice and Frequency

In the 1983 Morning Commuter Survey, individuals were asked different mode specific questions depending on their usual mode choice. The value which indicated the individual's usual mode choice was referred to as the "R-value". The R-value mode descriptions as defined by Hunt (1984) are:

.....walk all the way..............................R=1
.....use a private vehicle (car, van,
or pick-up) all the way.........
....as driver...............................R=2
....as a passenger........................R=3
....in a regular car-pool.................R=4
.....use bus all the way.......................R=5
.....use private vehicle part of the
way and then transfer to bus..............R=6
.....use LRT all the way.......................R=7
.....use a private vehicle part of the
way and then transfer to LRT..............R=8
.....use bus part of the way and then
transfer to LRT...........................R=9

The database for this study included car drivers (R=2), bus riders (R=5), LRT riders (R=7), and individuals who used bus part of the way and then transferred to LRT (R=9). Investigating the perceptions of other mode choice groups could be considered for possible future research.

Figure 4.4 contains the mode choice distribution for the database. This distribution likely contains a relatively larger number of public transit users making the morning home-to-work trip to the Edmonton CBD than exist in the Edmonton population. Again, this is because the CMCS data also exhibits this trend (Hunt, 1984).

For the cluster analysis procedure, two mode categories were considered: car drivers and public transit users. The mode categories names were as follows:

0 - car drivers
1 - public transport users

No assumptions were made regarding the perceptions of LRT riders. Investigations of the perceptions of LRT users as opposed to other public transit users could be examined in

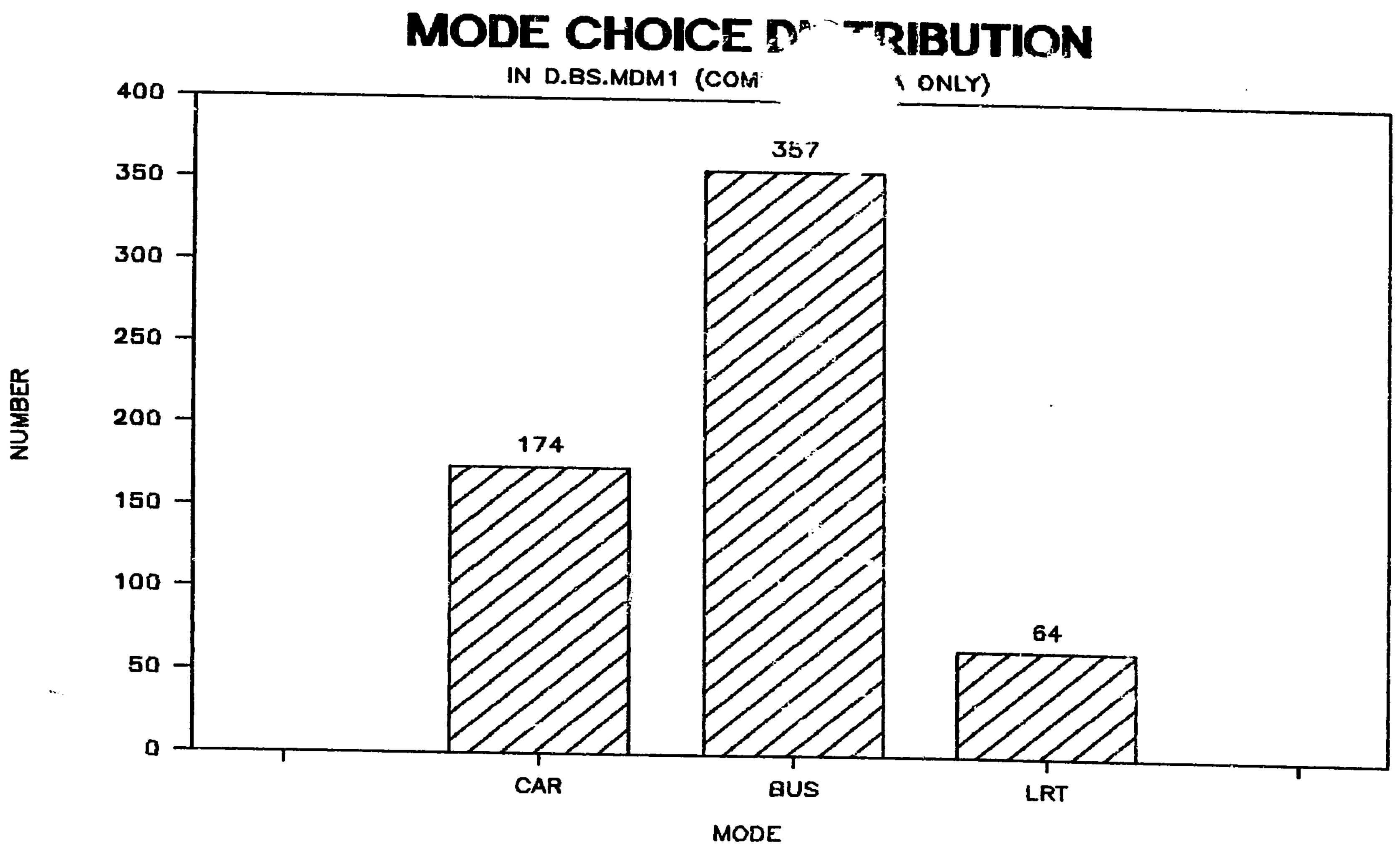


Figure 4.4 Mode Distribution in D.BS.MDM1 (Complete Data Only)

possible future research.

With regard to the mode frequency, each individual was asked how often he used his usual mode to make his home-to-work trip. All different answers to this question were coded as separate categories. The categories, and the number of individuals who belong to these categories are listed below (Hunt, 1984):

ANSWERS	NUMBER OF INDIVIDUALS
00 - no answer recorded	18
01 - "used every day", and "how often other modes used" question left blank	491
02 - "used every day", and "how often other modes used" question not left blank	10
03 - "used most days"	18
04 - "used most often"	27
05 - "today"	4
06 - "usual"	18
07 - "no alternative"	0
08 - "used regularly"	17
11 - "used almost everyday"	13
14 - "most"	0
24 - "quite often"	1
77 - if no answer recorded but "how often other modes used" question suggests "used most days"	18
78 - if no answer recorded but "how often other modes used" question suggests "today"	3
79 - if no answer recorded but "how often other modes used" question suggests "used most often"	0

It was felt that having so many categories which were almost the same, with so few members, would not give good cluster results. Therefore new categories were created by grouping synonymous categories together. The new categories were named such that the category value increased as the frequency of use increased. They are as follows:

CATEGORIES	NUMBER
0 - no answer recorded (00)	18
1 - "today" (05,78)	7
2 - "used most often" (04,24,)	28
3 - "used most days" (03,06,08,11,77)	84
4 - "used everyday" (01,02)	501

It is regrettable that numeric measures were not used to indicate how frequently individuals used their "usual mode" as the use of word categories introduces individuals' biases. If numeric values were given, the cluster procedure itself could have been used to determine the mode frequency groupings.

It should be noted, as is further discussed in Chapter 6, that upon completion of the cluster analysis it was discovered that only "regular" mode users were used in the logit model research conducted by Hunt (1984). Therefore, only "regular" users as were discerned by Hunt (1984) were included in the data set D.BS.MDM1 and are used in this research.

4.5 Variable Correlation Results

Before embarking on the cluster analysis, correlation analysis was employed to examine the linear relationships between the study variables.

Correlation coefficients, r_{ik}, provide a measure of linear association between two variables i and k (Johnson and Wichern, 1982). Their values range from +1 to -1, where +1 represents a strong tendency for positive linear association between two variables, and -1 represents a strong tendency for negative linear association between two variables. That is, if n=number of individuals and "a" and "b" are specific variables, a +1 correlation indicates a strong tendency for x_{ia} for $i \in n$ to get bigger as x_{ib} for $i \in n$ gets bigger and a -1 correlation indicates a strong tendency for the opposite (i.e. for x_{ia} for $i \in n$ to get smaller as x_{ib} for $i \in n$ gets bigger). If r=0, this implies that there is no linear relationship between the variables.

Using the Michigan Interactive Data Analysis System (MIDAS), Pearson Correlations between variable pairs, for all eleven variables described in this chapter, were calculated. The correlation results are illustrated in the matrix in Figure 4.5. Examination of this matrix reveals the following:

- as was expected, there are reasonably high correlations among the socioeconomic variables, i.e. between AGE, GENDER, JOB and INCOME
- as was expected, the perceived walk time variable, PWALK, is highly positively correlated with the actual walk distance, AWALK (R=0.5034)
- surprisingly, the perceived wait time variable, PWAIT, is not linearly correlated with the actual wait time variable, AWAIT. This implies that either (1) there is no relationship between these two variables, (2) the relationship that exists is not a linear one or (3) the actual wait time model in incorrect.

	PWALK	PWAIT	AWALK	AWAIT	COMPU	INCOME	AGE	GENDER	JOB	MODE	MODEFREQ
PWALK	1.0000										
PWAIT	0.2178	1.0000									
AWALK	0.5034	0.0169	1.0000								
AWAIT	-0.0234	0.0368	-0.1370	1.0000							
COMPU	-0.1612	-0.0812	-0.1574	-0.5530	1.0000						
INCOME	-0.0306	0.0671	0.0020	0.0624	-0.1489	1.0000					
AGE	0.0255	0.0268	-0.0364	0.0030	-0.0211	0.2710	1.0000				
GENDER	0.0529	-0.0251	-0.0044	-0.1499	0.2048	-0.5905	-0.1670	1.0000			
JOB	0.0412	-0.0615	0.0564	-0.0600	0.1029	-0.4903	-0.2316	0.2914	1.0000		
MODE	-0.1804	-0.3254	-0.0081	-0.1106	0.3005	-0.2064	-0.0413	0.1607	0.2489	1.0000	
MODEFREQ	-0.0061	0.0280	-0.0225	0.0265	0.0132	0.0323	0.1180	-0.0050	-0.0615	0.0604	1.0000

Figure 4.5 Pearson Correlation Matrix of Study Variables

the PWAIT variable is, however, somewhat positively correlated with the MODE variable. This indicates that perceived wait times tend to be higher for car drivers.

the composite utility variable, COMPU, is reasonably correlated with several of the variables (i.e. gender, mode, await). This implies that this is a good "explanatory" variable with regard to linearity.

the MODE variable is correlated with many of the variables, particularly, (1) negatively with the the INCOME variable which suggests higher income people tend to drive cars to work and (2) positively with the JOB variable which suggests that those persons with higher prestige jobs tend to drive to work.

5. FUZZY C-MEANS (FCM) PROGRAM RESULTS

The creation of the database enabled the cluster analysis stage of the research to begin. Program runs were made for fifteen different subsets of the input data. Descriptions of these different input data sets precedes the discussion of the run results. With regards to the run results, this chapter specifically addresses the following questions:

- For a particular number of clusters and set of input variables, given that solutions have been obtained from several different starting positions, what is the "best" solution?
- Having determined the "best" solutions for each number of clusters and combination of input variables investigated, which results best represent the cluster structure in the data?

5.1 Input Variable Values

This section will briefly describe the input variable values chosen for the FCM program runs. Table 5.1 summarizes the values for each progam run.

5.1.1 Scaling Matrix

Bezdek (1981) suggests that the scaling matrix chosen for a particular program run should correspond to the geometric and statistical properties of the data. Following is his description of data properties with which each scaling matrix is most compatible:

RUN	FILENAME	NUMBER OF DIMENSIONS	ICON	WEIGHTING EXPONENT	NUMBER OF CLUSTERS
1	OUT1.FUZZY1	11	1	2.00	2-5
2	OUT2.FUZZY1	11	2	2.00	2-5
3	OUT3.FUZZY2	11	3	2.00	2-5
4	OUT4.FUZZY2	11	3	2.00	6-9
5	OUT5.FUZZY2	11	3	2.00	15
6	OUT6.FUZZY2	11	3	2.00	25
7	OUT7.FUZZY2	6	3	2.00	2-5
8	OUT8.FUZZY2	6	3	1.25	2-5
9	OUT9.FUZZY2	[illegible]	3	1.25	2-5
10	OUT10.FUZZY2	[illegible]	3	1.25	2-5
11	OUT11.FUZZY2	9	3	1.25	2-5
12	OUT12.FUZZY2	10	3	1.25	2-5
13	OUT13.FUZZY2	9	3	1.25	6-9
14	OUT14.FUZZY2	10	3	1.25	6-9
15	OUT15.FUZZY2	11	3	1.25	6-9

Table 5.1 Description of Fuzzy C-Means Program Runs

SCALING MATRIX TYPE	DATA PROPERTIES
ICON=1 → $M=[I]$	-statistically independent, equally variable features for hyperspherical clusters
ICON=2 → $M=[\mathrm{diag}(\sigma_J^2)]^{-1}$	-statistically independent, unequally variable features for hyperellipsoidal clusters
ICON=3 → $M=[\mathrm{COV}(X)]^{-1}$	-statistically dependent, unequally variable features for hyperellipsoidal clusters

Correlation and variance-covariance matrices were produced to examine the statistical properties of the data. Both were calculated using MIDAS. All eleven variables described in Chapter 4 were included in the statistical analyses. Table 5.2 provides a summary of these variables with the abbreviations that will be used to describe them throughout the rest of the thesis. The variance-covariance results are illustrated in Figure 5.1, and the correlation results in Figure 4.5 (Section 4.5).

The variable variances are located on the diagonals of the variance-covariance matrix in Figure 5.1. Examination of these values indicates that large disparities exist among the variances. The values range from 0.1952 for the variance of the mode variable to 1.40E+06 for the variance of the age category variable.

The correlation matrix in Figure 4.5 indicates that several of the variables are highly correlated. For instance, as is expected, there is a reasonably high correlation between an individual's actual walk distance and

VARIABLE ABBREVIATION	DESCRIPTION
PWALK	- individual's perceived walk time
PWAIT	- individual's perceived wait time
AWALK	- individual's actual walk distance
AWAIT	- public transit service frequency
COMPU	- composite utility value
INCOME	- salary category of individual
AGE	- age category of individual
GENDER	- gender category of individual
JOB	- job category of individual
MODE	- individual's usual mode choice category
MODEFREQ	- category indicating the frequency of usual mode use

Table 5.2 Variable Abbreviations and Descriptions

	PWALK	PWAIT	AWALK	AWAIT	COMPU	INCOME	AGE	GENDER	JOB	MODE	MODEFREQ
PWALK	7.8824										
PWAIT	2.3970	15.360									
AWALK	262.89	12.314	34596.0								
AWAIT	-0.5068	1.1135	-196.93	59.688							
COMPU	-1.0939	-0.7697	-70.774	-10.329	5.8442						
INCOME	-0.1918	0.5872	0.8504	1.0762	-0.8041	4.9882					
AGE	84.671	124.36	-8016.5	27.178	-60.443	716.74	1.40E+06				
GENDER	0.0709	-0.0470	-0.3910	-0.5520	0.2360	-0.6287	-94.251	0.2273			
JOB	0.0727	-0.1515	6.5946	-0.2912	0.1563	-0.6982	-172.36	0.0873	0.3950		
MODE	0.2238	0.5634	0.6691	0.3776	-0.3209	0.2037	21.615	-0.0338	-0.0691	0.1952	
MODEFREQ	-0.0105	0.0672	-2.5642	0.1256	0.0197	0.0442	85.747	-0.0015	-0.0237	0.0164	0.3765

Figure 5.1 Matrix of Study Variables' Variances and Covariances

his perceived walk time. The correlation value, for these variables is 0.5034. As well, as is expected, correlations exist between the socioeconomic variables income, gender, age, and job type. Variable correlations are discussed in more detail in Section 4.5.

Therefore, because (1) the feature variances were unequal and (2) the variables were not statistically independent, the third scaling matrix type was considered to be the most suitable scaling matrix for this data. This choice of scaling matrix is indicated in the FCM program by an ICON value of 3 (for more information regarding the input variables see Section 3.5.2).

It should be noted that the ICON variable was not equal to 3 for the first two progam runs. Therefore, the results for these "trial" runs will not be studied in any further detail.

5.1.2 Weighting Exponent

At present, there exist no "rules" for determining the optimum value of the weighting exponent "s". In Section 3.3.4 it was suggested both that:

1. Values in-between 1.5 and 3.0 give good results (Bezdek, Ehrlich, and Full, 1984)
2. The value 2.0 is often used (Zimmerman, 1985)

As a starting point, the weighting exponent value was set equal to 2.0.

5.1.3 Number of Dimensions

For the first six program runs all 11 variables described in Chapter 4 were included. In later runs, the number of dimensions were varied, their number chosen largely because of the results from previous program runs. The dimensions chosen for these runs will be discussed in more detail when the program runs are examined.

5.2 General Description of Program Runs

5.2.1 Output Filenames

The naming scheme adopted for the output files listed in Table 5.1 was as follows: the output filename prefix indicated the run number, the suffix reflected whether the file was ouput from the FUZZY.1 or FUZZY.2 FCM program. Differences between these two programs are discussed in Section 3.5.7. All output files have been stored on magnetic tape filed in the University of Alberta magnetic tape library as volume 127505, rack number 071366.

Note that there is more than one output file associated with each set of input variables listed in Table 5.1, each derived from a different starting membership matrix ($U^{(0)}$). The membership matrix initialization process is discussed in more detail in Section 3.5.3. The output files in Table 5.1 contain results which have been calculated based on the membership initialization method which sets the RANDOM variable equal to 0.7731. Output files derived from initial membership matrices assigned by the RANDU subroutine are named such that the "root" name in Table 5.1 is retained along with a small letter extension at the name end, to distinguish runs from different random number generator seed numbers. For example, results obtained for the case where the value of the seed number was 1 contain the letter "a" at the filename end, for the next assumed seed number value the letter "b" was used. There are as many output files as was required to be confident that, for each number of clusters, for each set of input variables, optimum results were obtained. Detailed numbers and the exact process for determining optimum solutions are discussed next.

5.2.2 Optimum Program Results

As was discussed earlier, it is possible that more that one solution may result from different FCM program runs, where each solution has been obtained for a particular number of clusters, for a particular set of input data, by minimization of the objective function. The problem is to decide which of these solutions is the "best" solution. Simply choosing the results associated with the overall minimum objective function value may not be the most appropriate approach. Bezdek (1981) states that "objective functions usually have multiple local stationary points at fixed c, and global extrema are not necessarily the "best" c-partitions of the data" (Bezdek, 1981, p. 96). Cluster validity measures provide a means by which to compare different solutions. That is, results associated with the optimum cluster validity measures, minimum entropy and maximum partition coefficient values, may be deemed as best solutions.

It should be noted that these values do not necessarily correspond to the minimum objective function values. An example of this can be seen in the OUT8.FUZZY2 run summary in Figure 5.7. Here, for the four cluster case, the minimum entropy value of 0.363 and the maximum partition coefficient value of 0.805 occur in solution 1. For this same run the objective function is equal to 0.195E+04, the maximum objective function value attained for this number of clusters. The best solution, then, as identified by the cluster validity measures is solution 1; not the solution suggested as optimum by the objective function.

Bezdek (1974) outlines the approach he used to choose the best solution from a set of solutions generated from different initial matrices. He used only the partition coefficient to gauge how good the cluster structure was for a particular solution. The solution which was associated with the maximum partition coefficient, for a particular number of clusters, was deemed as the best solution for that number of clusters. This was done for all number of clusters considered. It was these best solutions that were later investigated to determine what number of clusters best described the cluster structure.

For this research an approach similar to Bezdek's (1974) was used. Here, however, the entropy measure rather than the partition coefficient was used to gauge the results. This measure was used because, according to Bezdek, Ehrlich, and Full (1984), it is slightly more sensitive to partition quality than is the partition coefficient. The strategy used to determine the optimum results was then as follows. For a given set of input data, for each number of clusters specified in the cluster range, results were generated from two different initial membership matrices one from the process which assigns the RANDOM variable equal to 0.7731 and the other from the process which calls the RANDU subroutine using a seed number of 1. If, for the specific number of clusters, the entropy values were the same, no more program runs were made. Solutions were considered the same when the entropy values were within 0.005 of each other. It is interesting to note that Bezdek (1973) considered solutions to be the same when the partition coefficients were within 0.01 of each other. If the entropy values were different, more runs were made using different seed numbers (variable IX in the FCM Program) in the RANDU subroutine, until either the results associated with the minimum entropy value were duplicated, or five progam runs were made. The seed numbers used for runs 2 through 5 were 1, 111111, 215, and 43561, respectively. If after 5 runs, the results associated with the minimum entropy value had not been duplicated, the minimum entropy value results were accepted as the optimum results. This occurrence could imply either that the surface over which the objective function is being minimized was very hilly or, perhaps, the program was stagnating at a local minimum. It is recognized that it may be somewhat risky to label results that only have been duplicated twice, or minimum entropy results that have not been duplicated in 5 runs as optimum results; however, because of limits on program running costs, these assumptions are considered to be reasonable.

The FCM program results for the last thirteen sets of input variables (described in Table 5.1) are summarized in Figures 5.2 through 5.14. In each summary the minimum entropy values have been underlined, to indicate which run results, for each number of clusters, are considered "best" results.

5.2.3 Convergence Characteristics

Bezdek, Ehrlich and Full (1984) make the following statement with regard to the FCM program's convergence:

> "Practically speaking, no difficulties have ever been encountered, and numerical convergence is usually achieved in 10-25 iterations" (Bezdek, Ehrlich, and Full, 1984, p. 194)

Examination of the run summaries in Figures 5.2 through 5.14, particularly for the runs summarized in Figures 5.7 through 5.14, reveals that for this data the number of iterations required for convergence was high. Numbers of

OUT3.FUZZY2 RESULTS

NUMBER OF VARIABLES = 11
SCALING MATRIX TYPE = 3
WEIGHTING EXPONENT = 2.00

PARTITION COEFFICIENT

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.500	0.500	-	-	-
3	0.333	0.333	-	-	-
4	0.250	0.250	-	-	-
5	0.200	0.200	-	-	-

ENTROPY MEASURE

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.693	0.693	-	-	-
3	1.098	1.098	-	-	-
4	1.386	1.386	-	-	-
5	1.609	1.609	-	-	-

OBJECTIVE FUNCTION VALUE (X10000)

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.293	0.293	-	-	-
3	0.195	0.195	-	-	-
4	0.146	0.146	-	-	-
5	0.117	0.117	-	-	-

NUMBER OF ITERATIONS

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	7	3	-	-	-
3	7	4	-	-	-
4	7	3	-	-	-
5	6	3	-	-	-

Figure 5.2 OUT3.FUZZY2 Solution Summary

OUT4.FUZZY2 RESULTS

NUMBER OF VARIABLES = 11
SCALING MATRIX TYPE = 3
WEIGHTING EXPONENT = 2.00

NUMBER OF	PARTITION COEFFICIENT				
CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
6	0.167	0.167	-	-	-
7	0.143	0.143	-	-	-
8	0.125	0.125	-	-	-
9	0.111	0.111	-	-	-

NUMBER OF	ENTROPY MEASURE				
CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
6	1.791	1.791	-	-	-
7	1.945	1.945	-	-	-
8	2.078	2.078	-	-	-
9	2.196	2.196	-	-	-

NUMBER OF	OBJECTIVE FUNCTION VALUE (X10000)				
CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
6	0.098	0.098	-	-	-
7	0.084	0.084	-	-	-
8	0.073	0.073	-	-	-
9	0.065	0.065	-	-	-

NUMBER OF	NUMBER OF ITERATIONS				
CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
6	6	3	-	-	-
7	6	3	-	-	-
8	6	3	-	-	-
9	6	2	-	-	-

Figure 5.3 OUT4.FUZZY2 Solution Summary

OUT5.FUZZY2 RESULTS

NUMBER OF VARIABLES = 11
SCALING MATRIX TYPE = 3
WEIGHTING EXPONENT = 2.00

	PARTITION COEFFICIENT				
NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
15	0.067	0.067	-	-	-

	ENTROPY MEASURE				
NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
15	2.706	2.706	-	-	-

	OBJECTIVE FUNCTION VALUE (X10000)				
NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
15	0.039	0.039	-	-	-

	NUMBER OF ITERATIONS				
NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
15	5	3	-	-	-

Figure 5.4 OUT5.FUZZY2 Solution Summary

OUT6.FUZZY2 RESULTS

NUMBER OF VARIABLES = 11
SCALING MATRIX TYPE = 3
WEIGHTING EXPONENT = 2.00

PARTITION COEFFICIENT

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
25	0.063	0.062	-	-	-

ENTROPY MEASURE

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
25	3.740	3.738	-	-	-

OBJECTIVE FUNCTION VALUE (X10000)

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
25	0.479	0.483	-	-	-

NUMBER OF ITERATIONS

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
25	5	3	-	-	-

Figure 5.5 OUT6.FUZZY2 Solution Summary

OUT7.FUZZY2 RESULTS

NUMBER OF VARIABLES = 6
SCALING MATRIX TYPE = 3
WEIGHTING EXPONENT = 2.00

PARTITION COEFFICIENT

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.502	0.500	-	-	-
3	0.336	0.333	-	-	-
4	0.252	0.250	-	-	-
5	0.202	0.200	-	-	-

ENTROPY MEASURE

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.691	0.693	-	-	-
3	1.095	1.098	-	-	-
4	1.383	1.386	-	-	-
5	1.605	1.608	-	-	-

OBJECTIVE FUNCTION VALUE (X10000)

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.160	0.160	-	-	-
3	0.106	0.106	-	-	-
4	0.080	0.080	-	-	-
5	0.064	0.064	-	-	-

NUMBER OF ITERATIONS

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	14	4	-	-	-
3	14	4	-	-	-
4	14	3	-	-	-
5	14	3	-	-	-

Figure 5.6 OUT7.FUZZY2 Solution Summary

OUT8.FUZZY2 RESULTS

NUMBER OF VARIABLES = 6
SCALING MATRIX TYPE = 3
WEIGHTING EXPONENT = 2.00

PARTITION COEFFICIENT

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.905	0.904	-	-	-
3	0.802	0.802	-	-	-
4	0.805	0.794	0.794	0.794	0.794
5	0.806	0.805	-	-	-

ENTROPY MEASURE

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.171	0.171	-	-	-
3	0.352	0.353	-	-	-
4	0.363	0.391	0.391	0.391	0.391
5	0.380	0.385	-	-	-

OBJECTIVE FUNCTION VALUE (X10000)

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.252	0.252	-	-	-
3	0.216	0.216	-	-	-
4	0.195	0.192	0.192	0.192	0.192
5	0.174	0.176	-	-	-

NUMBER OF ITERATIONS

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	10	10	-	-	-
3	18	46	-	-	-
4	13	19	21	20	18
5	19	26	-	-	-

Figure 5.7 OUT8.FUZZY2 Solution Summary

OUT9.FUZZY2 RESULTS

NUMBER OF VARIABLES = 11
SCALING MATRIX TYPE = 3
WEIGHTING EXPONENT = 1.25

PARTITION COEFFICIENT

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.626	0.626	-	-	-
3	0.569	0.570	-	-	-
4	0.580	0.581	-	-	-
5	0.599	0.616	0.599	0.615	-

ENTROPY MEASURE

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.552	0.552	-	-	-
3	0.748	0.748	-	-	-
4	0.807	0.804	-	-	-
5	0.821	0.783	0.821	0.785	-

OBJECTIVE FUNCTION VALUE (X10000)

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.490	0.490	-	-	-
3	0.440	0.439	-	-	-
4	0.404	0.404	-	-	-
5	0.376	0.375	0.376	0.376	-

NUMBER OF ITERATIONS

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	20	23	-	-	-
3	25	21	-	-	-
4	20	42	-	-	-
5	25	40	43	36	-

Figure 5.8 OUT9.FUZZY2 Solution Summary

OUT10.FUZZY2 RESULTS

NUMBER OF VARIABLES = 8
SCALING MATRIX TYPE = 3
WEIGHTING EXPONENT = 1.25

PARTITION COEFFICIENT

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.791	0.791	-	-	-
3	0.780	0.676	0.780	-	-
4	0.707	0.708	-	-	-
5	0.707	0.680	0.696	0.702	0.696

ENTROPY MEASURE

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.340	0.340	-	-	-
3	0.415	0.571	0.414	-	-
4	0.562	0.560	-	-	-
5	0.587	0.634	0.613	0.606	0.613

OBJECTIVE FUNCTION VALUE (X10000)

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.349	0.349	-	-	-
3	0.306	0.310	0.306	-	-
4	0.278	0.278	-	-	-
5	0.259	0.259	0.258	0.260	0.258

NUMBER OF ITERATIONS

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	32	17	-	-	-
3	21	16	16	-	-
4	47	24	-	-	-
5	16	37	25	29	27

Figure 5.9 OUT10.FUZZY2 Solution Summary

OUT11.FUZZY2 RESULTS

NUMBER OF VARIABLES = 9
SCALING MATRIX TYPE = 3
WEIGHTING EXPONENT = 1.25

PARTITION COEFFICIENT

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.729	0.728	-	-	-
3	0.698	0.673	0.673	0.674	0.698
4	0.692	0.691	-	-	-
5	0.686	0.657	0.647	0.648	0.659

ENTROPY MEASURE

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.426	0.426	-	-	-
3	0.549	0.584	0.585	0.584	0.549
4	0.604	0.604	-	-	-
5	0.649	0.697	0.709	0.708	0.693

OBJECTIVE FUNCTION VALUE (X10000)

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.397	0.397	-	-	-
3	0.352	0.352	0.352	0.352	0.352
4	0.319	0.319	-	-	-
5	0.300	0.299	0.297	0.297	0.298

NUMBER OF ITERATIONS

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	26	20	-	-	-
3	56	16	17	17	56
4	33	20	-	-	-
5	43	46	24	28	26

Figure 5.10 OUT11.FUZZY2 Solution Summary

OUT12.FUZZY2 RESULTS

NUMBER OF VARIABLES = 10
SCALING MATRIX TYPE = 3
WEIGHTING EXPONENT = 1.25

PARTITION COEFFICIENT

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.676	0.675	-	-	-
3	0.623	0.624	-	-	-
4	0.660	0.638	0.639	0.638	0.638
5	0.671	0.671	-	-	-

ENTROPY MEASURE

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.490	0.491	-	-	-
3	0.661	0.662	-	-	-
4	0.658	0.700	0.699	0.700	0.700
5	0.672	0.672	-	-	-

OBJECTIVE FUNCTION VALUE (X10000)

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	0.444	0.444	-	-	-
3	0.397	0.396	-	-	-
4	0.363	0.362	0.362	0.362	0.362
5	0.334	0.334	-	-	-

NUMBER OF ITERATIONS

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
2	24	23	-	-	-
3	20	18	-	-	-
4	32	29	26	23	34
5	57	43	-	-	-

Figure 5.11 OUT12.FUZZY2 Solution Summary

OUT13.FUZZY2 RESULTS

NUMBER OF VARIABLES = 9
SCALING MATRIX TYPE = 3
WEIGHTING EXPONENT = 1.25

PARTITION COEFFICIENT

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
6	0.653	0.636	0.645	0.636	0.645
7	0.652	0.644	0.653	0.644	0.659

ENTROPY MEASURE

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
6	0.722	0.754	0.740	0.755	0.740
7	0.737	0.759	0.749	0.759	0.725

OBJECTIVE FUNCTION VALUE (X10000)

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
6	0.281	0.280	0.280	0.280	0.280
7	0.265	0.265	0.266	0.265	0.265

NUMBER OF ITERATIONS

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
6	22	28	31	27	31
7	34	34	32	46	62

Figure 5.12 OUT13.FUZZY2 Solution Summary

OUT14.FUZZY2 RESULTS

NUMBER OF VARIABLES = 10
SCALING MATRIX TYPE = 3
WEIGHTING EXPONENT = 1.25

NUMBER OF	PARTITION COEFFICIENT				
CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
6	0.640	0.640	-	-	-
7	0.634	0.641	0.641	-	-

NUMBER OF	ENTROPY MEASURE				
CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
6	0.752	0.752	-	-	-
7	0.787	0.772	0.772	-	-

NUMBER OF	OBJECTIVE FUNCTION VALUE (X10000)				
CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
6	0.313	0.313	-	-	-
7	0.298	0.297	0.297	-	-

NUMBER OF	NUMBER OF ITERATIONS				
CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
6	26	26	-	-	-
7	33	33	34	-	-

Figure 5.13 OUT14.FUZZY2 Solution Summary

OUT15.FUZZY2 RESULTS

NUMBER OF VARIABLES = 11
SCALING MATRIX TYPE = 3
WEIGHTING EXPONENT = 1.25

PARTITION COEFFICIENT

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
6	0.577	0.639	0.576	0.576	0.639
7	0.598	0.599	-	-	-

ENTROPY MEASURE

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
6	0.880	0.771	0.882	0.882	0.771
7	0.864	0.864	-	-	-

OBJECTIVE FUNCTION VALUE (X10000)

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
6	0.354	0.351	0.354	0.354	0.351
7	0.333	0.333	-	-	-

NUMBER OF ITERATIONS

NUMBER OF CLUSTERS	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
6	29	32	33	29	34
7	35	32	-	-	-

Figure 5.14 OUT15.FUZZY2 Solution Summary

iterations greater than 25 were reasonably common, and on three occasions the maximum number of iterations specified internally in the FCM program had to be increased beyond the previously specified maximum of 50 (see Section 3.5.4.). These occurred for:

1. OUT11.FUZZY2, Solution 5, 3 clusters
2. OUT12.FUZZY2, Solution 1, 5 clusters
3. OUT13.FUZZY2, Solution 5, 7 clusters

For these cases the maximum number of iterations was reset to 75 iterations, and the program was rerun.

The large number of iterations required for convergence of the FCM program, for these data sets, may have been a result of:

1. The large number of dimensions.
2. The characteristics of the data itself. Perhaps, the characteristics of the data were such that the surface over which the objective function was being minimized was very hilly and, therefore, many iterations were required for the algorithm's convergence.

5.2.4 Program Running Time

The amount of program running time required to run the FCM program was found to be very much a function of the number of iterations required for the Fuzzy Cluster algorithm's convergence. For this reason it would be very hard to estimate how much time would be required for a given program run. Table 5.3 lists the program running times for the last thirteen sets of input data which were obtained using the RANDU subroutine with a root value of 1.0. The last three run times are unavailable because the running time limit of 400 cpu seconds was exceeded. For these program runs the cluster range was specified for 6 to 9 clusters; however, in 400 cpu seconds results were only obtained for the 6 and 7 cluster cases. These results proved to be enough for the analysis of the program runs and, therefore, the results for the 8 and 9 cluster solutions were not required.

Generally speaking, the program running time seems to increase as the number of dimensions increases. It also seems to increase as the number of clusters increases. As an aside, it is interesting to note that Bezdek, Ehrlich, and Full (1984) state the the number of iterations generally increases if the third scaling matrix is used. This was not investigated in this research.

5.3 Program Results

5.3.1 Runs 1 and 2

For reasons provided earlier, results from these two program runs are considered invalid and, therefore, have not been examined in any further detail.

FILENAME	NUMBER OF DIMENSIONS	NUMBER OF CLUSTERS	PROGRAM RUN TIME (SECONDS)
OUT3.FUZZY2	11	2-5	26.0
OUT4.FUZZY2	11	6-9	70.1
OUT5.FUZZY2	11	15	65.8
OUT6.FUZZY2	11	25	168.9
OUT7.FUZZY2	6	2-5	11.7
OUT8.FUZZY2	6	2-5	17.3
OUT9.FUZZY2	11	2-5	123.2
OUT10.FUZZY2	8	2-5	122.8
OUT11.FUZZY2	9	2-5	163.4
OUT12.FUZZY2	10	2-5	201.7
OUT13.FUZZY2	9	6-9	N/A
OUT14.FUZZY2	10	6-9	N/A
OUT15.FUZZY2	11	6-9	N/A

Table 5.3 Fuzzy C-Means Program Run Times

5.3.2 Runs 3,4,5, and 6

The intention of these program runs was to investigate what effect varying the number of clusters had on the cluster results. For each run the ICON variable was set to 3 and the weighting exponent to 2.0. All 11 dimensions described in Chapter 4 were used. The cluster range varied between runs, as is illustrated in Table 5.1.

The "best solution" entropy values for each number of clusters, are plotted on a graph shown in Figure 5.15. Also plotted on this graph is the ln(c) function. To interpret the meaning of these results recall:

1. The objective is to minimize the entropy value
2. The maximum entropy value possible, for a particular number of clusters, is equal to ln(c).

Upon examination of the Entropy Trend graph it appears that this second statement is violated. For the case where the number of clusters is 25, the entropy value lies above the ln(c) curve. Because this entropy value was derived twice (see Figure 5.5), this result is attributed to round-off error. It is suspected that this entropy value should plot along the ln(c) curve with all of the other entropy values.

Having all entropy values plot along the ln(c) curve implies that there is no cluster structure in the data, essentially, that all objects belong equally to all clusters. Examination of the partition coefficient values in Figure 5.16 yields the same results. Here, where the objective is to maximize the partition coefficient, all

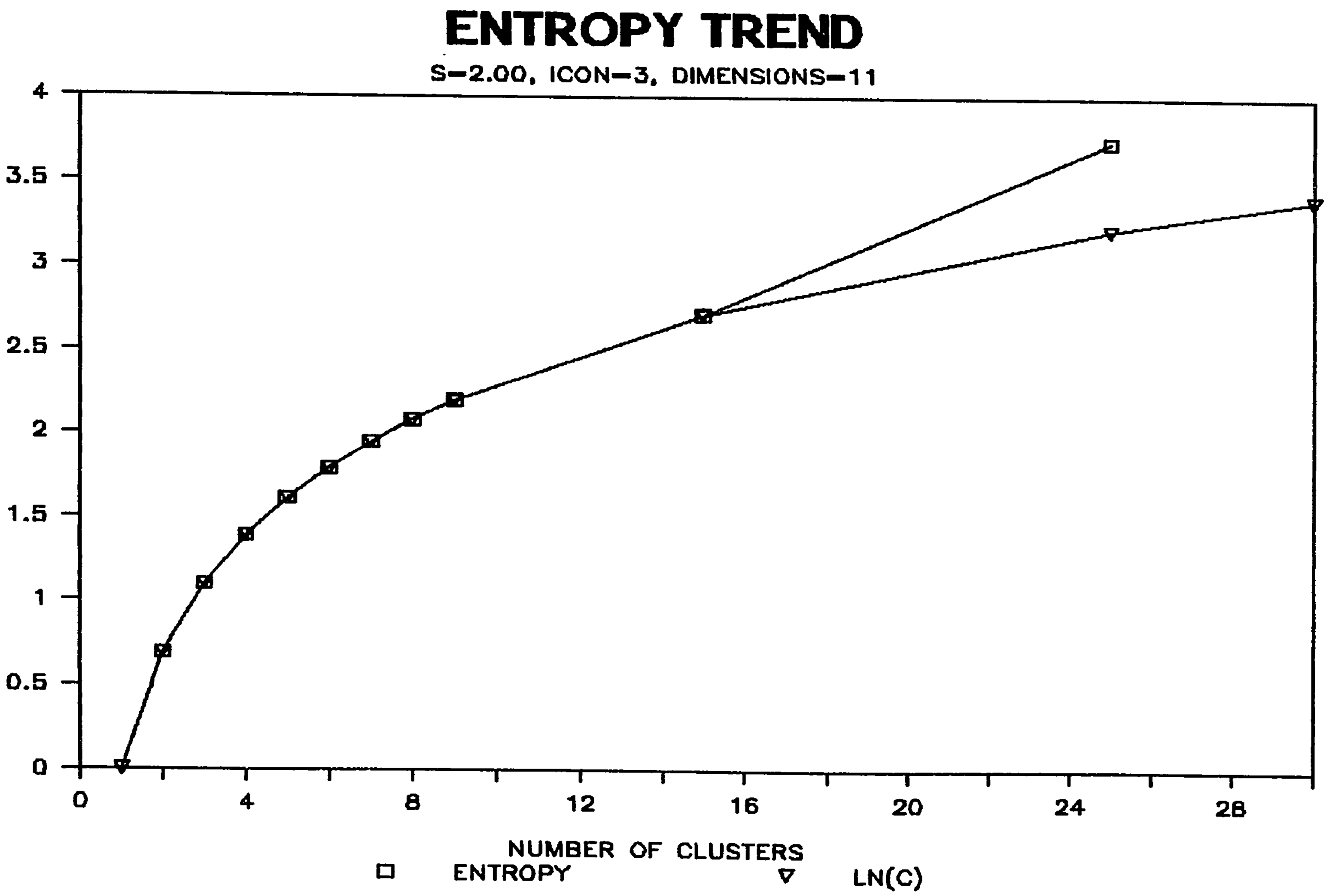


Figure 5.15 Entropy Trend for FCM Program Runs 3 through 6

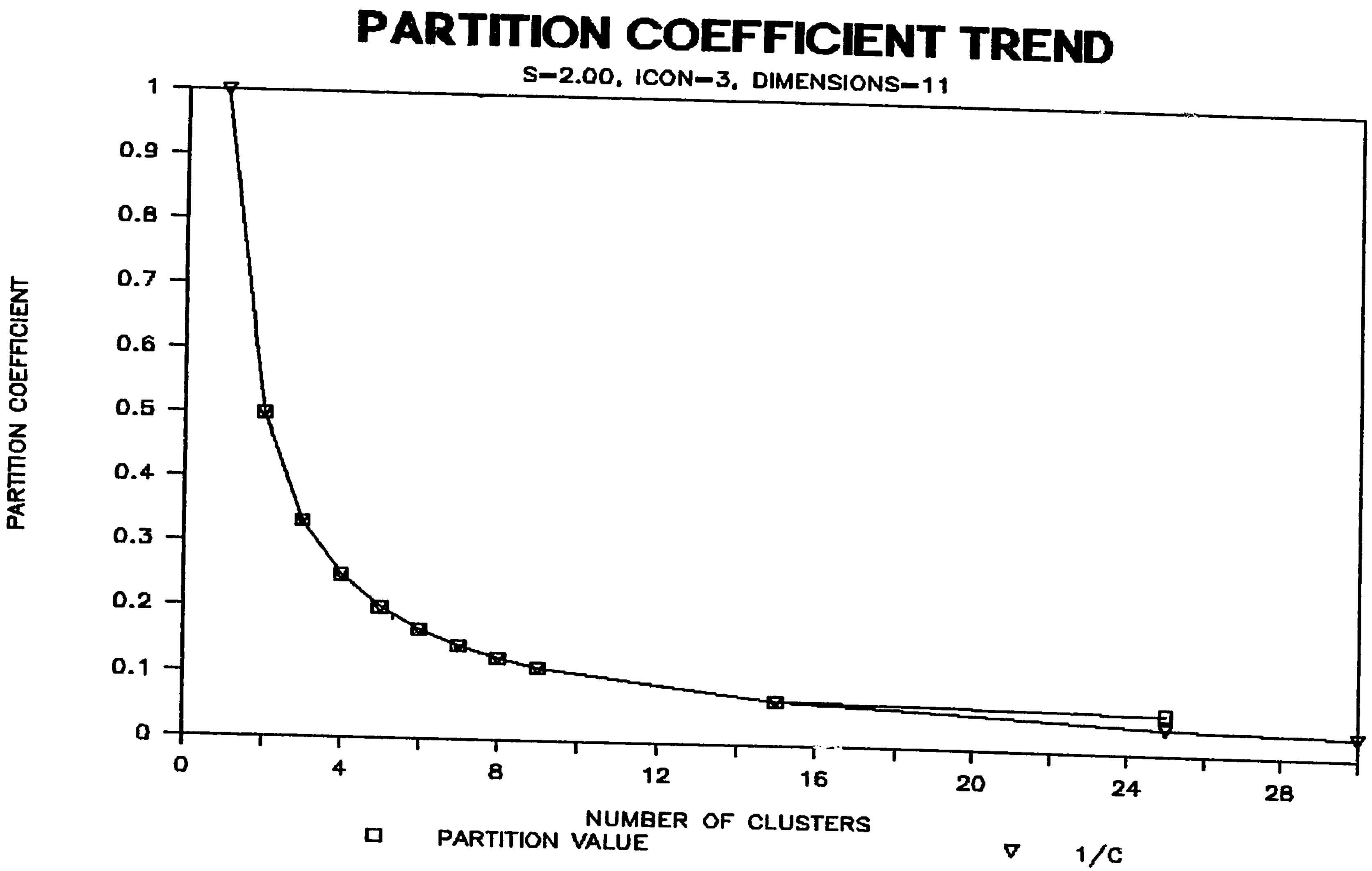


Figure 5.16 Partition Coefficient Trend for FCM Program Runs 3 through 6

values plot along the $\frac{1}{c}$ curve, the minimum possible partition coefficient values for a given number of clusters.

Initially, it was felt that there were, perhaps, too many dimensions being considered for reasonable results to occur. Therefore, for the next program runs the number of dimensions was reduced.

5.3.3 Runs 7 and 8

Since past research suggested that car users' and public transport users' perceptions of public transport attributes were in fact different, the perceived and actual public transport variables as well as the mode variable were chosen as the input variables for run 7. The composite utility variable was also included because of its good explanatory power as evidenced by the correlation matrix (Section 4.5). For run 7 the ICON value and weighting exponent were not changed from the values used in runs 3 through 6. A cluster range of 2-5 clusters was specified. The "best solution" entropy values were plotted on a graph shown in Figure 5.17. Again, the entropy results indicated that no cluster structure existed in the data.

Next, the effect of adjusting the weighting exponent was investigated. According to Bezdek (1981) "the larger s is, the "fuzzier" are the membership assignments; and conversely, as $s \overset{+}{\to} 1$, fuzzy c-means solutions become hard" (Bezdek, 1981, p. 70). Since what was desired was a "harder" solution the value of s was reduced to 1.25. Again, the

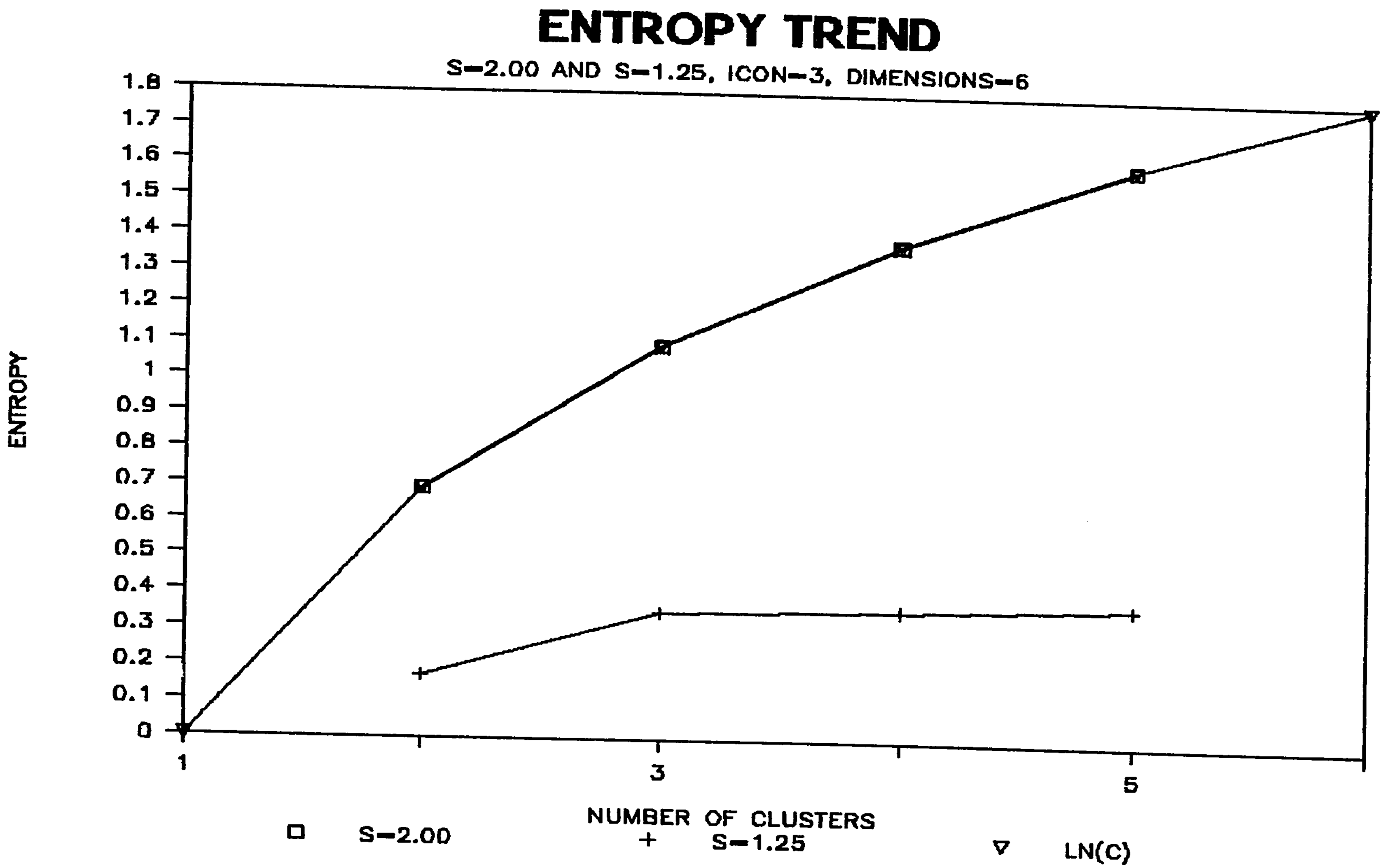


Figure 5.17 Entropy Trend for FCM Program Runs 7 and 8

choice of s was a rather arbitrary one with the only constraint being that its value be greater that one (Bezdek, Ehrlich, and Full, 1984). Therefore a weighting exponent of 1.25 and the same scaling matrix, 6 dimensions and the cluster range described above for run 7 were used as the input variables for run 8. The entropy results for this run are also plotted on the graph in Figure 5.17.

The graph indicates that the reduction of the weighting exponent value did result in a cluster structure being assigned to the data. It is not completely understood what this means with regards to the data. One possible explanation is that the clusters identified by the algorithm are not well-separated and consequently, when the higher weighting exponent was used all objects were assigned equally to the specified number of clusters. When the weighting exponent was then reduced, the program was more-or-less forced into crisper membership assignments. Because of the imprecision associated with perceptions, one might expect to see not-so-well-separated clusters be developed. In fact, development of clusters of this nature was justification for the use of the Fuzzy Cluster Method. The fact that no cluster structure was identified when the weighting exponent was equal to 2.0 will be considered when the cluster results are further analysed.

Because of practical limits on time and program running costs, and because this research is primarily concerned with testing the applicability of the Fuzzy Cluster Method for the investigation of travelers' perceptions, the effect of varying the weighting exponent was not investigated any further. This, however, requires more research in the future.

5.3.4 Runs 9,10,11, and 12

For runs 9,10,11, and 12 the weighting exponent was set equal to 1.25, the ICON value to 3, and only the number of dimensions was varied. It was felt that it was desirable to use as many explanatory variables that provided good cluster results so that the information that could be derived from the cluster results was maximized.

The decision of which variables to add as the number of dimensions was increased was made partly on the results from the correlation matrix (Figure 4.5) with some subjective judgements. Basically, variables were added based on their correlation with the perceived variables: highly correlated variables being added first, less correlated variables being added later. Figure 5.18 summarizes the dimensions considered in each of the program runs.

The entropy results for these runs are plotted in Figure 5.19; the partition coefficient results in Figure 5.20. Both plots indicate that there is some cluster structure in the data.

Having determined that cluster structure existed in the data, the next step was to determine what combination of variables and number of clusters best described this cluster

OUTPUT	NUMBER OF	DIMENSIONS										
FILENAME	DIMENSIONS	PWALK	PWAIT	AWALK	AWAIT	COMPU	MODE	INCOME	GENDER	AGE	JOB	MODEFREQ
OUT8.FUZZY2	6	X	X	X	X	X	X					
OUT9.FUZZY2	11	X	X	X	X	X	X	X	X	X	X	X
OUT10.FUZZY2	8	X	X	X	X	X	X	X	X			
OUT11.FUZZY2	9	X	X	X	X	X	X	X	X	X		
OUT12.FUZZY2	10	X	X	X	X	X	X	X	X	X	X	

Figure 5.18 Matrix of Variables Considered in FCM Program Runs 9,10,11 and 12

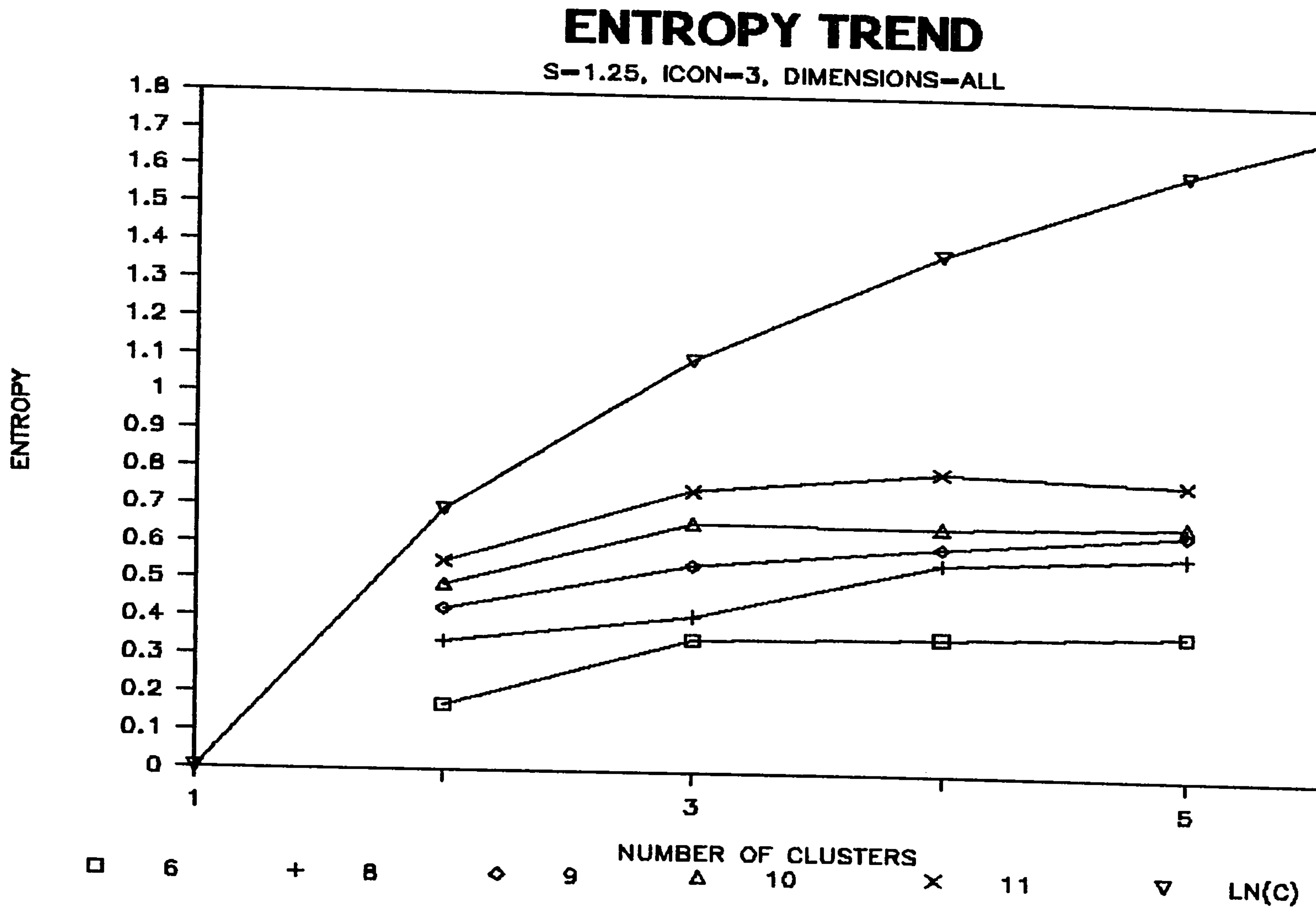


Figure 5.19 Entropy Trend for FCM Program Runs 9 through 12

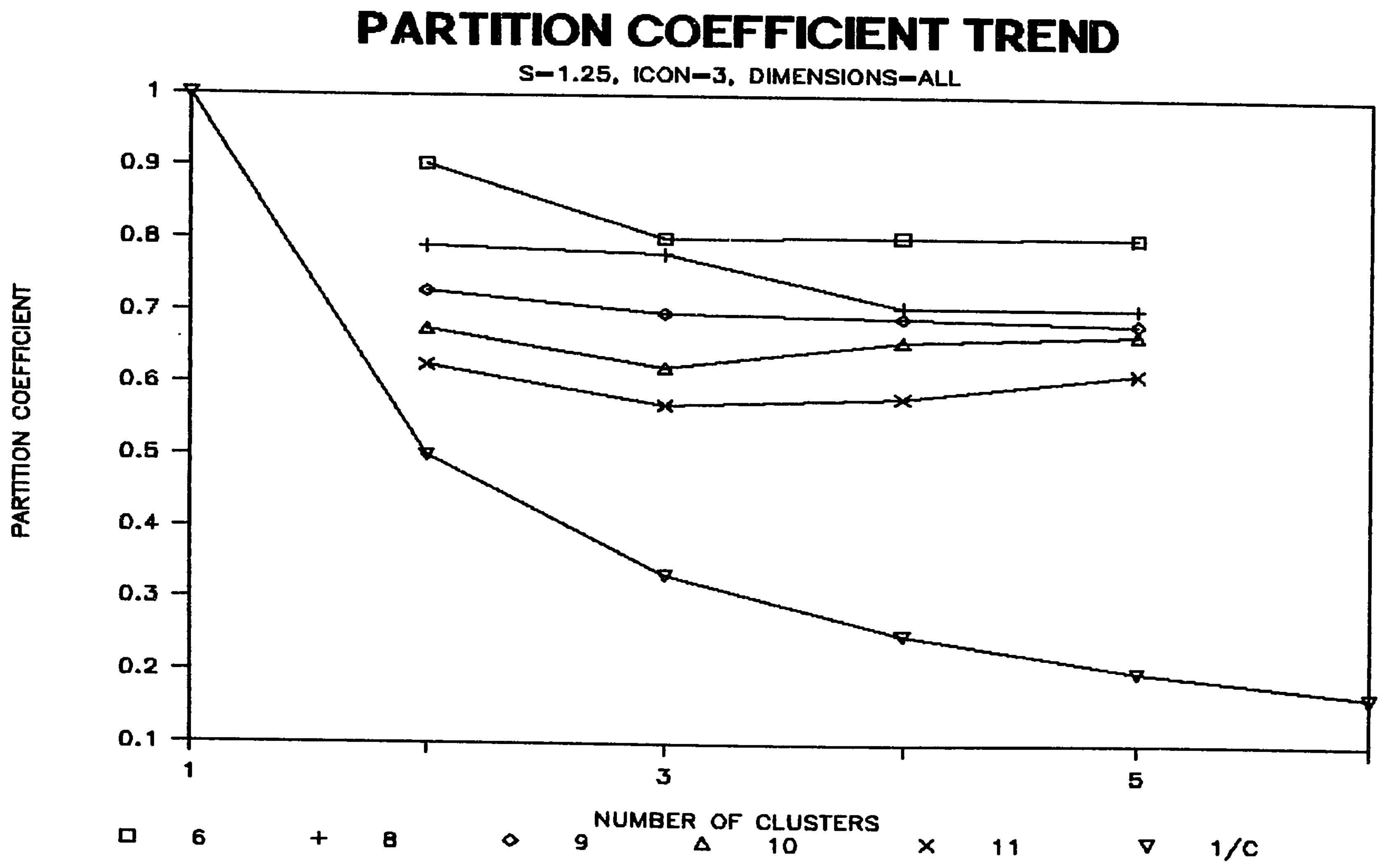


Figure 5.20 Partition Coefficient Trend for FCM Program Runs 9 through 12

structure. As was stated in Chapter 3.4, one of the problems with both the entropy measure and the partition coefficient is that they tend to indicate that the most valid number of clusters is two. Certainly, this seemed to be the case for these results. Based on entropy minimization, the entropy values in Figure 5.19 always indicated that the best number of clusters was two, regardless of the number of dimensions being considered. This same result, based on maximization of the partition coefficient, was evident from the partition coefficient trend graph (Figure 5.20).

For cases such as this an alternative method of analysis was outlined in Chapter 3.4. What was suggested was that the trends of the cluster validity measures be studied, rather than simply choosing the results associated with either the maximum or minimum of a particular cluster validity measure as the "best" results. Therefore, for the entropy measure the optimum number of clusters would be that number associated with the entropy value which lies below the trend of entropy values for the other number of clusters. The opposite would be true for the optimum partition coefficient. It would lie above the trend of the partition coefficient values for the other number of clusters (Zimmerman, 1985).

Windham (1981) suggests this same approach be used for the analysis of the "best" number of clusters. He states that "entropy tends to increase with c independent of structure in the data. So, the value of c indicated by entropy as optimum, is the one for which entropy falls below an increasing trend in its values" (Windham, 1981, p. 183). He illustrates this idea in a graph which has been included in this thesis as Figure 5.21. He suggests that the partition coefficient be analyzed in a similar manner, i.e. the manner already described above.

It is worth noting that not only does the entropy value increase as the number of clusters increases, but it also seems to increase as the number of dimensions increases (Figure 5.19). Intuitively this seems to be reasonable. One would expect that as the information provided by the number of dimensions is increased, it would become more difficult for an object to possess all the properties required to belong to a particular cluster. Therefore, on average, the membership values would tend to be fuzzier, and the entropy values associated with the cluster structure would increase.

This trend also has implications for the analysis of the "best" clusters. That is, the results that are associated with the absolute minimum entropy value will not necessarily be regarded as the "best" results. Rather, the results for each number of dimensions will be analysed independently to determine what number of clusters best describes the cluster structure. This analysis is described in Chapter 6.

Using the method described above, the Entropy Trend graph in Figure 5.19 was examined to determine the "best" number of clusters. For the 9,10, and 11 dimension results

FIGURE OMITTED

DUE TO

COPYRIGHT

RESTRICTION

Figure 5.21 Optimum Number of Clusters from Examination of Entropy Trend (Windham, 1981, p. 183)

the entropy trend was not completely evident and, therefore, before the best cluster stucture problem could be addressed the cluster range for these dimensions had to be increased. The last three program runs were made for this purpose.

5.3.5 Runs 13,14, and 15

These program runs were made for the 9,10 and 11 dimension cases with a cluster range of 6 to 9 clusters. As was discussed previously, results were only obtained for the 6 and 7 cluster cases.

The entropy results for these number of clusters were added to those determined for the 2 to 5 cluster range and plotted in Figure 5.22. The partition coefficient results were plotted in Figure 5.23.

With the entropy trends now evident for these dimensions, the question of what number of clusters best described the cluster structure could be addressed.

5.4 Best Results

Using the method outlined by Windham (1981), the Entropy Trend graphs in Figures 5.19 and 5.22 were examined to determine the "best" number of clusters. The following were deemed as solutions which best represented the cluster structure for the data, for the specific dimensions considered.

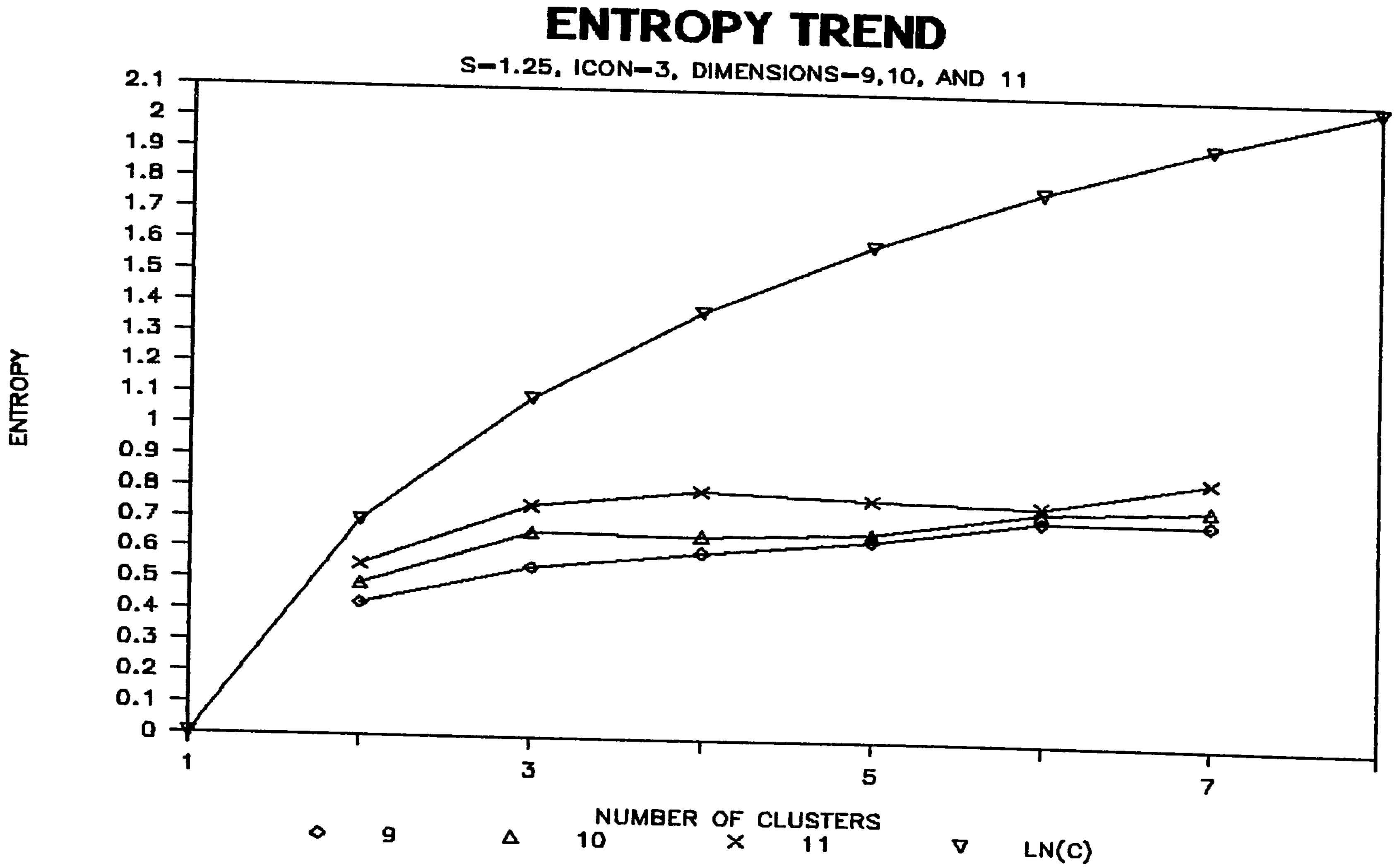


Figure 5.22 Entropy Trend for FCM Program Runs 13, 14 and 15

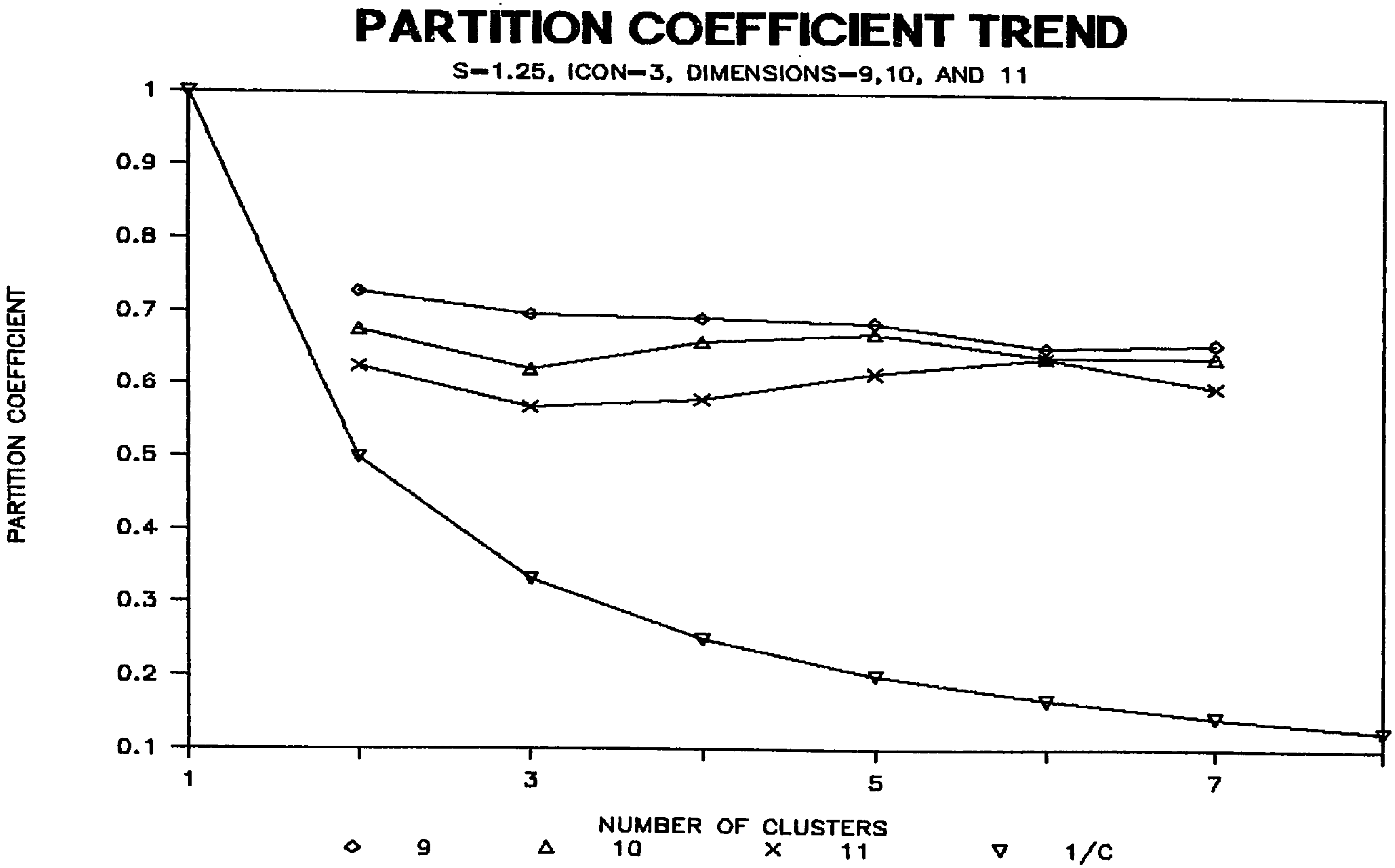


Figure 5.23 Partition Coefficient Trend for FCM Program Runs 13, 14, and 15

NUMBER OF DIMENSIONS	"BEST" NUMBER OF CLUSTERS
6	2
8	3
10	4,5
11	6

The following observations were made regarding the choice of these "best" number of clusters:

1. No entropy value seemed to fall below the trend for the 9 dimension entropy plot. For this reason no "best" number of clusters was specified.
2. For the 10 dimensional case, the entropy values associated with both the 4 and 5 cluster solutions seemed to lie below the entropy trend. Therefore, both of these solutions were considered as "best" results.

The "best" results predicted by the partition coefficient trend in Figures 5.20 and 5.23 agree with those predicted by the entropy measure.

Analysis of the characteristics of the clusters that have been identified in these "best" solutions is the topic of the next chapter.

6. ANALYSIS OF THE "BEST" CLUSTER RESULTS

Chapter 6 examines the characteristics of the "best" cluster structures identified in Chapter 5. Two methods were employed for this task. The first method used the cluster centre values as a means of comparing the average cluster characteristics; the second examined the characteristics of the cluster "core" and used these cores as a means of comparison.

6.1 Description of Cluster Properties Using Cluster Centre Values

Using cluster centre values for examination of the cluster characteristics assumes that these values represent the average characteristic values for the cluster members. To supplement this assumption, it is desirable to measure the standard deviation of the cluster characteristics and then, ideally, to test whether the cluster centre values differ significantly, statistically, from one another. There is, however, to the knowledge of the author, no measure of standard deviation for the cluster centre values and development of such a statistic is beyond the scope of this research.

In the absence of a standard deviation value, two physical characteristics of the cluster were measured. These characteristics were the cluster size and fuzziness. Fuzziness is a reasonably good proxy for the standard deviation value as it also measures deviation; the deviation of membership values in the cluster.

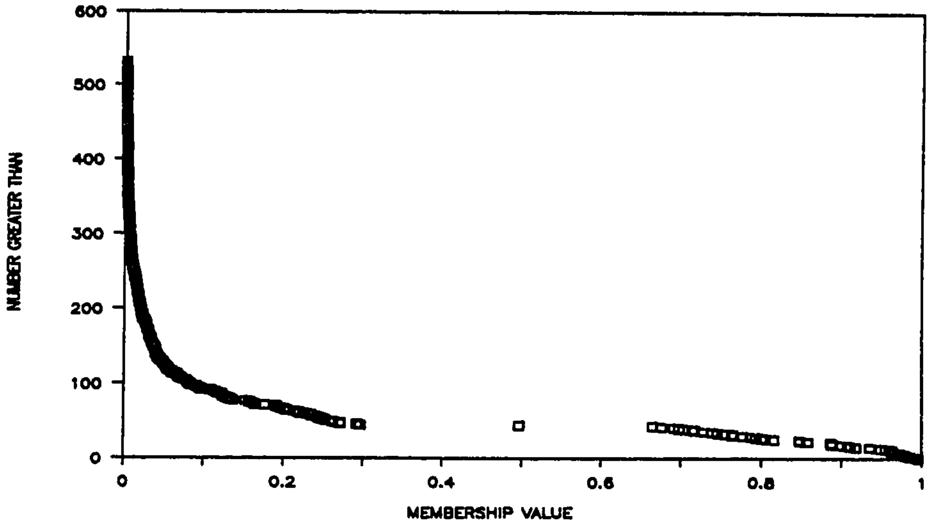
6.1.1 Cluster Size

Because an object's membership into a cluster is defined by a degree of membership, the number of objects which belong to a cluster, or the traditional size of the cluster, is not clearly defined. Two methods for evaluating cluster size were developed for this research. One method examines cluster size graphically; the other provides a numerical measure which is later used in evaluating cluster fuzziness.

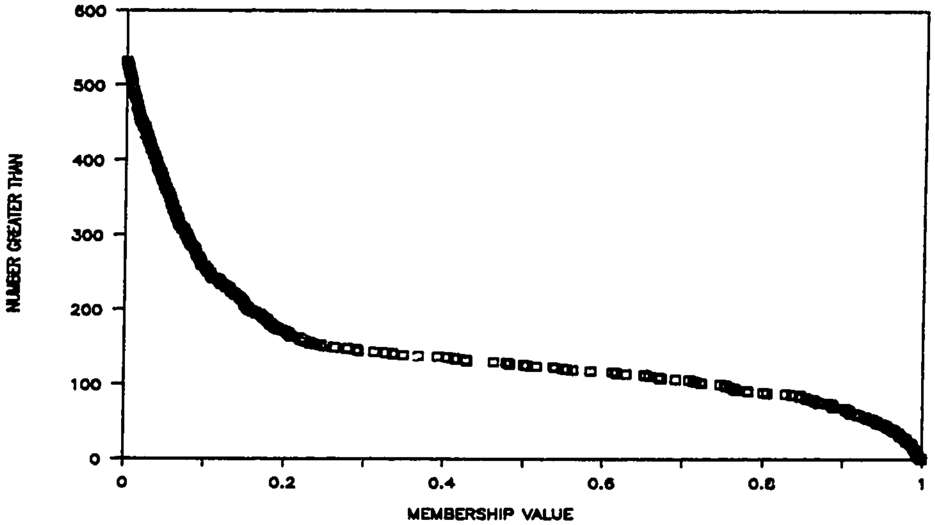
Graphically, cluster size was examined by plotting the cumulative distribution of object membership values for the cluster. Figure 6.1 illustrates cumulative membership distribution graphs for clusters 5 and 6, from the six cluster solution developed in Chapter 5. In each graph, objects' membership values are plotted on the X-axis and the corresponding number of objects whose membership is greater than these values are plotted on the Y-axis. This is done for all objects in the sample data.

It is assumed that "larger" clusters contain a greater number of objects with high membership values. Using this logic the sizes of clusters from the same solution may be compared. The graphs in Figure 6.1 indicate that cluster 6 is larger than cluster 5.

The numerical size measure for fuzzy clusters was based on the crisp cluster size definition. In crisp set theory



Cluster 5 Membership Distribution



Cluster 6 Membership Distribution

Figure 6.1 Membership Distribution Graphs for Clusters 5 and 6

objects' membership values are either one or zero, depending on whether they belong to the set or not. Cluster size, then, may be considered to be the number of objects which belong to the cluster, or because of the one and zero memberships, the total summation of all objects' memberships into the cluster. Similarly, the numeric measure used to define cluster size for fuzzy clusters can be defined as the summation of object memberships for the cluster whose size is being investigated. Mathematically;

$$\text{CLUSTER SIZE}_i=\sum_{k=1}^{n}\mu_{ik}$$

Note that the size of all clusters together, or the total sum of all clusters' membership values equals the number of objects in the sample (for this research 532).

The cluster size measure has been included in the upper right hand corner of the cluster membership graphs.

6.1.2 Cluster Fuzziness

Bezdek's (1981) Partition Entropy was the basis for the cluster fuzziness measure. Recall that Partition Entropy measures the fuzziness of the cluster solution itself. Its formulation is:

$$H(U;C)=-\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}\ln(\mu_{ik})$$

where $\mu_{ik}\ln(\mu_{ik})=0$ whenever $\mu_{ik}=0$. Essentially, this measure is the summation of entropy values ($\mu_{ik}\ln(\mu_{ik})$) for all objects, into all clusters, normalized by the number of objects in the sample.

An analagous cluster entropy measure would be the summation of entropy values for all objects belonging to the cluster normalized by the number of objects in the cluster. The sum of membership values for a particular cluster could represent the number of objects in the cluster. Mathematically, then, the entropy for a particular cluster would be defined as:

$$HCE(U)_i = -\frac{1}{\sum_{k=1}^{n}\mu_{ik}} \sum_{k=1}^{n}\mu_{ik}\ln(\mu_{ik})$$

In this research, this measure was referred to as the Cluster Entropy measure. As is characteristic of the Partition Entropy measure, less fuzzy results will be reflected by smaller Cluster Entropy values.

Because this measure was used for comparing the fuzziness of clusters which were developed from the same solution, further investigation into the mathematical properties of this measure were deemed unnecessary and beyond the scope of this research. The Cluster Entropy value is also included in the upper right hand corner of the membership distribution graphs.

The membership distribution graphs also provide information regarding a cluster's fuzziness. Crisp clusters are characterized by membership graphs whose membership values are concentrated at the tail ends of the graphs (at the high or low membership regions). In contrast, membership graphs associated with fuzzier clusters contain more membership values in-between these two regions. By definition, fuzzy clusters contain a large proportion of memberships in the $\frac{1}{c}$ membership region. On this basis, cluster 6 from Figure 6.1 would be deemed "fuzzier" than cluster 5.

6.1.3 Perception Measures

Except for the perception values, the cluster centre values were used to compare and contrast cluster characteristics between clusters. For the comparison of the perceptions values it was felt that the measure of comparison should also reflect the corresponding actual variable values. Therefore, two perception measures were used; the first was the ratio of the perceived and actual cluster centre travel time values, and the second was the difference between the actual and perceived values. Mathematically:

1. Perception Ratio = $\frac{\text{Perceived Travel Time Value}}{\text{Actual Travel Time Value}}$
2. Perception Difference = (Actual - Perceived) Travel Time Value

During the analysis it was realized that the definitions of the measures were somewhat confusing, as individuals who underestimated their travel time values had smaller Perception Ratio values and larger Perception Difference values than those individuals who overestimated their travel times. In retrospect, these measures should have been defined such that small or large values for each represented the same result.

Actual walk time and wait time values were required for calculation of these measures. These times were estimated assuming:

1. An average walking speed of 1.2 m/s.
2. A public transport wait time equal to one half of the public transit service frequency (see Section 4.3.1).

The errors in these assumptions are discussed below.

First, the assumption that all individuals walk at a constant walking speed is incorrect because it does not account for such factors as walking environment, i.e. residential walking speeds versus downtown walking speeds impeded by stops at signals etc., the age or gender of the individual, etc. These factors likely cause variations in walking speeds and present errors in the walk time calculation. Second, the walking speed assumption of 1.2 m/s is likely too low. According to Nicholson (1987), walking speeds for Edmonton pedestrians range from 1.39 to 1.71 m/s. The assumption of a 1.2 m/w walking speed, then, likely results in actual walk time estimations that are too high.

The assumed wait time model was a rather crude wait time estimate. The model was used because, to the knowledge of the author, there existed no better wait time model for Edmonton commuters at the time of the research.

The effects of these errors are considered to be negligible as the same assumptions were used for all individuals in the sample, and the results for which they were the basis were used for comparative purposes only.

6.2 Analysis of Results Using Cluster Centre Values

The cluster centre values, the cluster perception characteristics, the cluster size and fuzziness were examined for each "best" solution described in Chapter 5. The chart in Figure 6.2 highlights these results. Tabulated summaries of this information and corresponding cluster membership distribution graphs are contained in Appendix B.

Following is a detailed examination of the characteristics of the "best" solution clusters from the chart in Figure 6.2.

6.2.1 Two Cluster Results

The sample data, in its aggregate form, is represented at the top of the chart in Figure 6.2. Below this are the characteristics for the two clusters defined in the "best" two cluster solution developed in Chapter 5 (highlighted in Figure 6.3). Following is a description of the two cluster result with Table 6.1 illustrating the cluster solution summary format and Figure 6.4 illustrating the cumulative membership distribution graph format from Appendix B.

Cluster 1 is made up almost entirely of public transit users, cluster 2 almost entirely of car users (MODE values of 0.98 and 0.09, respectively). Examination of the perception measures for these clusters reveals:

TWO CLUSTER SUMMARY

DIMENSION	CLUSTER	
	1	2
PWALK	3.344	4.259
PWAIT	3.885	6.197
AWALK	253.063	287.488
AWAIT	12.396	14.377
COMPU	-7.052	-8.561
MODE	0.979	0.090

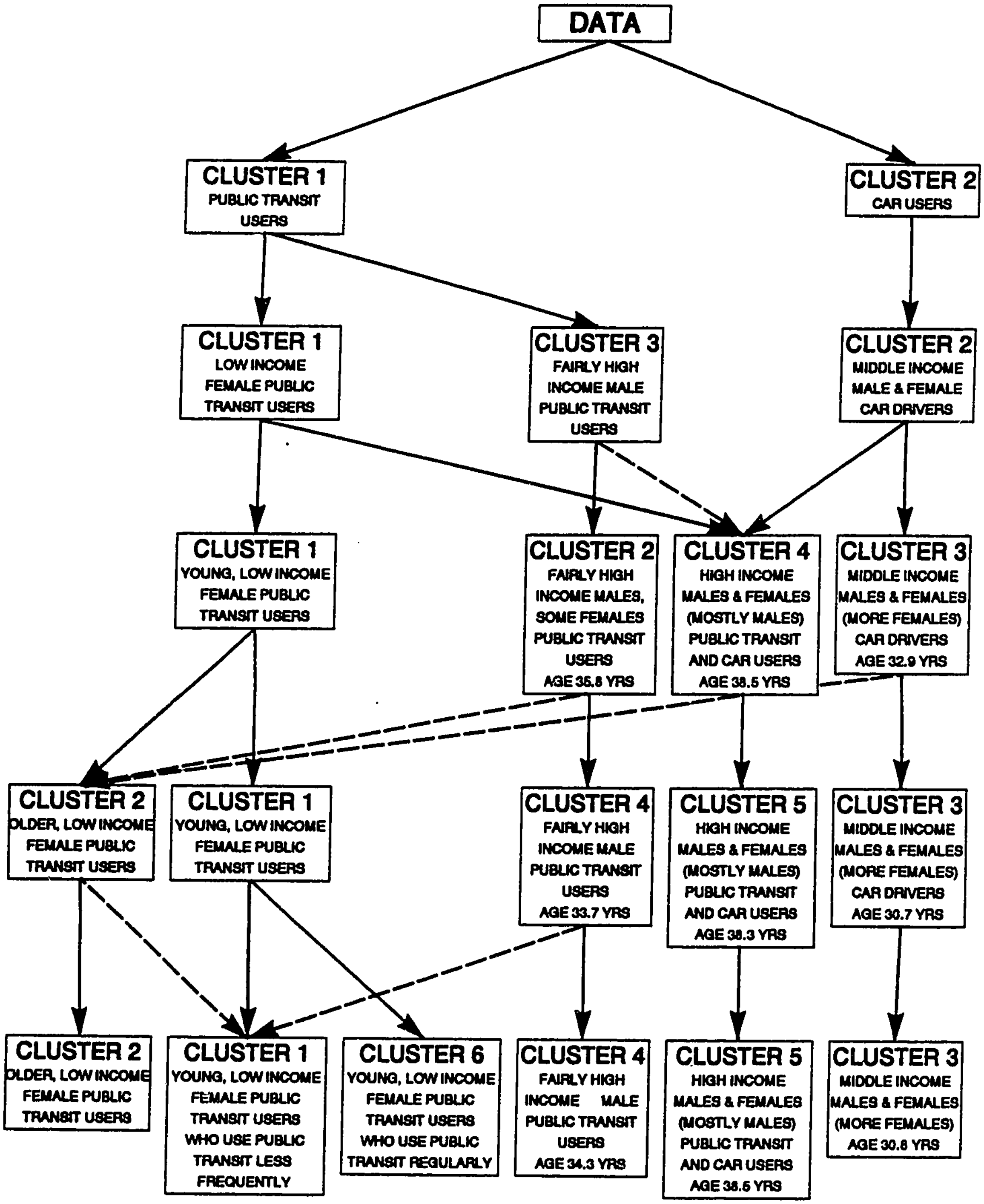
PERCEPTION RATIOS (PERCEIVED/ACTUAL):

WALK	0.95	1.08
WAIT	0.63	0.86

PERCEPTION DIFFERENCES (ACTUAL - PERCEIVED):

WALK (min)	0.17	-0.27
WAIT (min)	2.31	0.99

Table 6.1 Summary of Cluster Centre Characteristics for Two Cluster Solution


DATA
CLUSTER 1 PUBLIC TRANSIT USERS
CLUSTER 2 CAR USERS
CLUSTER 1 LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 3 FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS
CLUSTER 2 MIDDLE INCOME MALE & FEMALE CAR DRIVERS
CLUSTER 1 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 2 FAIRLY HIGH INCOME MALES, SOME FEMALES PUBLIC TRANSIT USERS AGE 35.8 YRS
CLUSTER 4 HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS AGE 38.5 YRS
CLUSTER 3 MIDDLE INCOME MALES & FEMALES (MORE FEMALES) CAR DRIVERS AGE 32.9 YRS
CLUSTER 2 OLDER, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 1 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 4 FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS AGE 33.7 YRS
CLUSTER 5 HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS AGE 38.3 YRS
CLUSTER 3 MIDDLE INCOME MALES & FEMALES (MORE FEMALES) CAR DRIVERS AGE 30.7 YRS
CLUSTER 2 OLDER, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 1 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS WHO USE PUBLIC TRANSIT LESS FREQUENTLY
CLUSTER 6 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS WHO USE PUBLIC TRANSIT REGULARLY
CLUSTER 4 FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS AGE 34.3 YRS
CLUSTER 5 HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS AGE 38.5 YRS
CLUSTER 3 MIDDLE INCOME MALES & FEMALES (MORE FEMALES) AGE 30.6 YRS

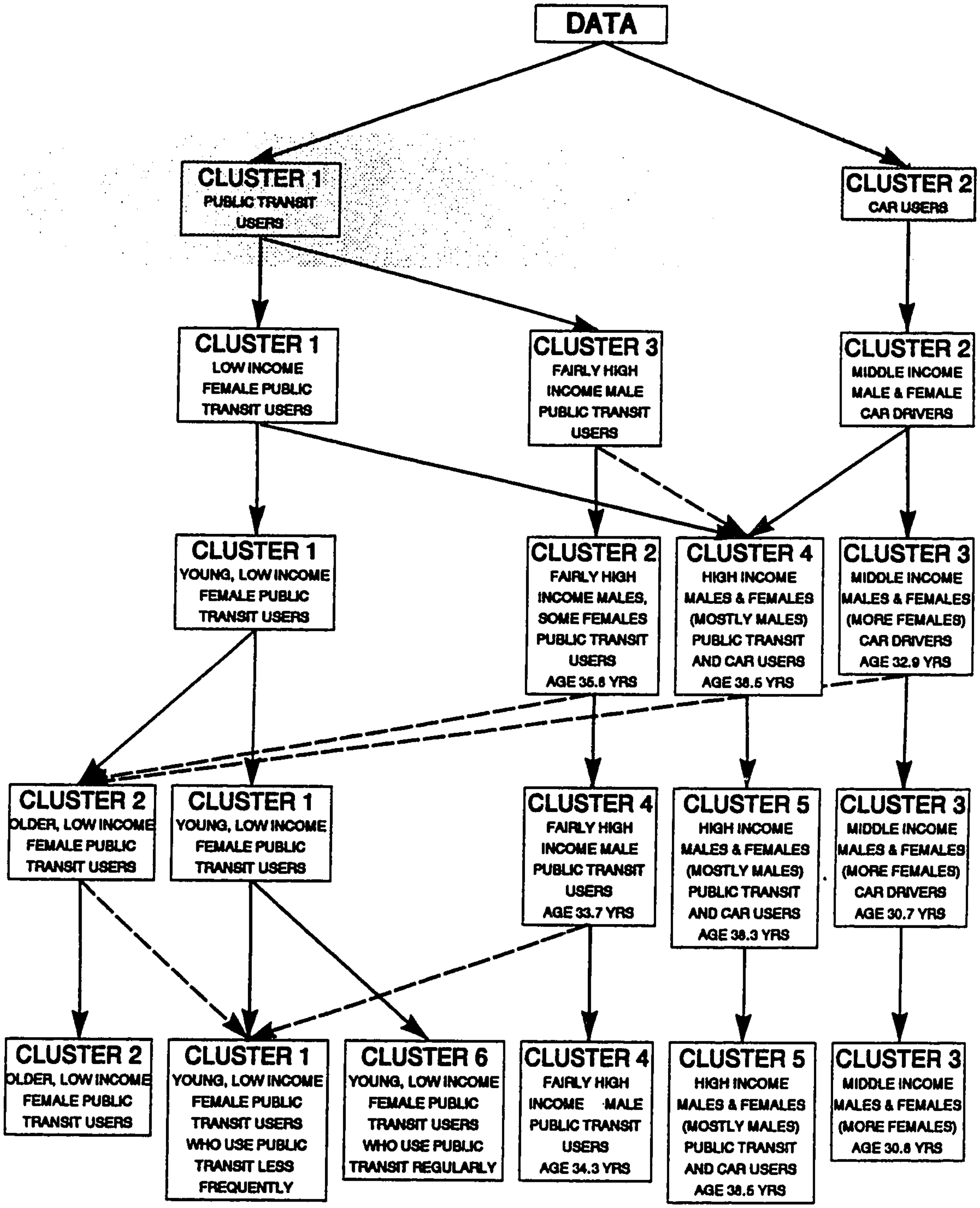

DATA
CLUSTER 1 PUBLIC TRANSIT USERS
CLUSTER 2 CAR USERS
CLUSTER 1 LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 3 FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS
CLUSTER 2 MIDDLE INCOME MALE & FEMALE CAR DRIVERS
CLUSTER 1 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 2 FAIRLY HIGH INCOME MALES, SOME FEMALES PUBLIC TRANSIT USERS AGE 35.8 YRS
CLUSTER 4 HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS AGE 38.5 YRS
CLUSTER 3 MIDDLE INCOME MALES & FEMALES (MORE FEMALES) CAR DRIVERS AGE 32.9 YRS
CLUSTER 2 OLDER, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 1 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 4 FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS AGE 33.7 YRS
CLUSTER 5 HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS AGE 38.3 YRS
CLUSTER 3 MIDDLE INCOME MALES & FEMALES (MORE FEMALES) CAR DRIVERS AGE 30.7 YRS
CLUSTER 2 OLDER, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 1 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS WHO USE PUBLIC TRANSIT LESS FREQUENTLY
CLUSTER 6 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS WHO USE PUBLIC TRANSIT REGULARLY
CLUSTER 4 FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS AGE 34.3 YRS
CLUSTER 5 HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS AGE 38.5 YRS
CLUSTER 3 MIDDLE INCOME MALES & FEMALES (MORE FEMALES) AGE 30.8 YRS

Figure 6.3 Cluster Characteristics from the Two Cluster Solution

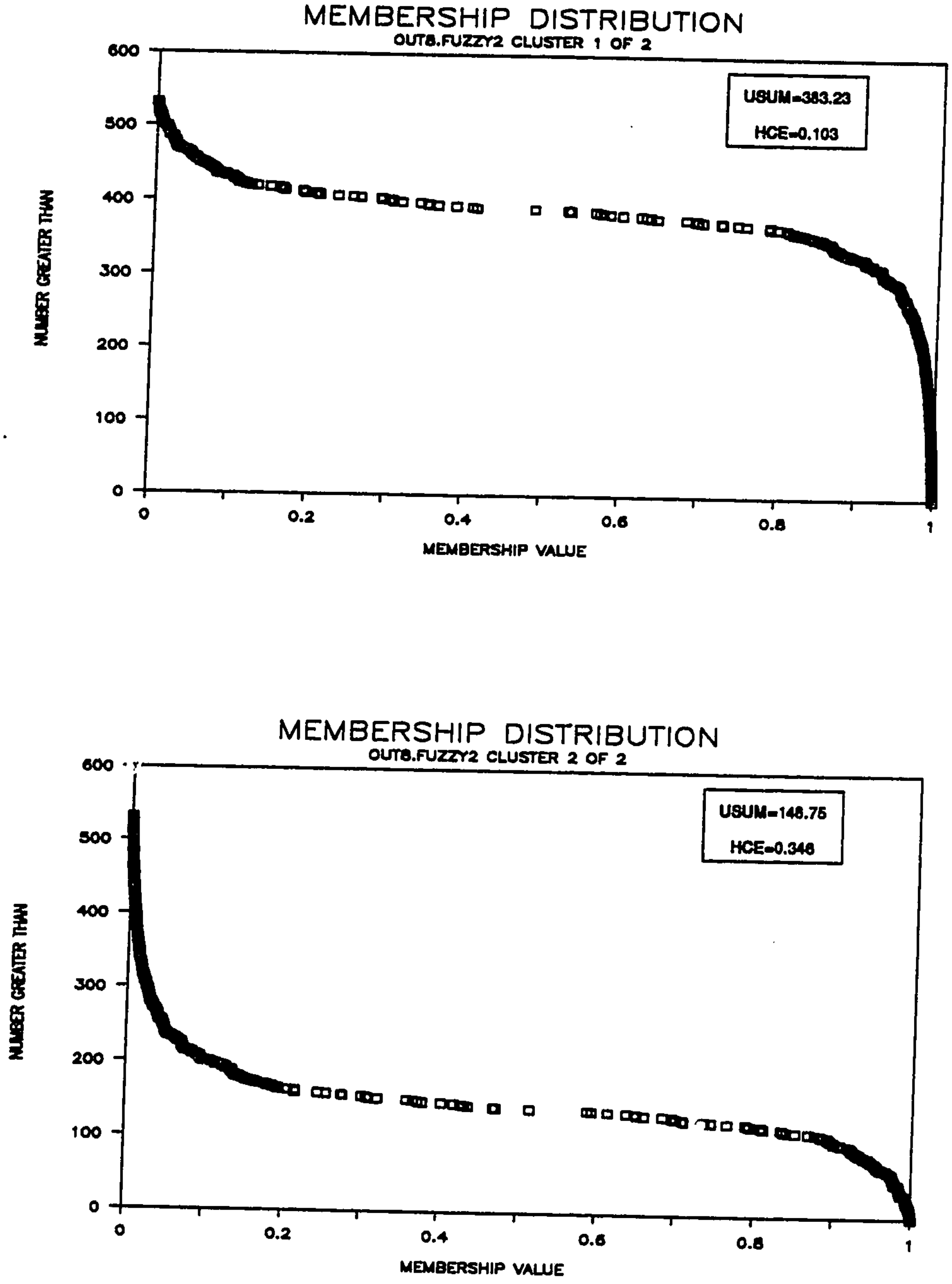


Figure 6.4 Cumulative Membership Distribution Graphs for Clusters 1 and 2 for the Two Cluster Solution

1. On average, the public transport users' perceptions of public transit walk times are less than those of the private auto drivers.
2. On average, the public transport users' perceptions of public transit wait times are less than those of the private auto drivers.

These results reflect the findings in the literature. (Quarmby, 1967; O'Farrell and Markham, 1974; Heggie, 1976; Meyburg and Brog, 1981). It is interesting to note that the difference between the walk time perceptions are less than the wait time values. Examination of the walk time difference values reveal that, on average, those individuals who belong to cluster 1 underestimate the public transit walk time by 10.2 seconds (.17 minutes), and those who belong to cluster 2 overestimate this time by 16.0 seconds (-.27 minutes). Both estimates appear to be reasonably accurate. On the other hand, those individuals from cluster 1 underestimate the public transport wait time, on average, by 2.3 minutes and those from cluster two underestimate this time by 1.0 minute. The difference between these two estimations is 1.32 minutes, a substantially larger amount.

Examination of the cluster membership distribution graphs (Figure 6.4) indicates that cluster 1 is the larger of the two clusters. The cluster size values indicate the same result. Because of the large proportion of public transit users in the data sample, this is a logical result.

The membership distribution graphs indicate that these two clusters are relatively "crisp" clusters. There is a heavy concentration of membership values at the tails of each graph and relatively few membership values in the in-between region. The Cluster Entropy values suggest that cluster 1 is less fuzzy than cluster 2.

6.2.2 Three Cluster Results

With the addition of the gender and age variables, three clusters were found to best represent the cluster structure in the data. Figure 6.5 highlights the "three cluster level" from Figure 6.2.

As was the case for the two cluster solution, clusters have again been divided on the basis of usual mode choice. Clusters 1 and 3 contain, almost entirely, public transit users and cluster 2 automobile users. Differences between clusters 1 and 3 are of a socioeconomic nature: Cluster 1 consists largely of females who earn, on average, 16,250 $/year, cluster 3 consists mostly of males who earn, on average, 28,700 $/yr. Cluster 2 contains almost an equal amount of males and females (GENDER=1.56). This is a middle income group of individuals who earn approximately 25,350 $/yr.

As would be expected, the walk and wait time estimations for the public transit users, clusters 1 and 3, are smaller than those for the car users of cluster 2. It is interesting to note that the perceptions for the higher

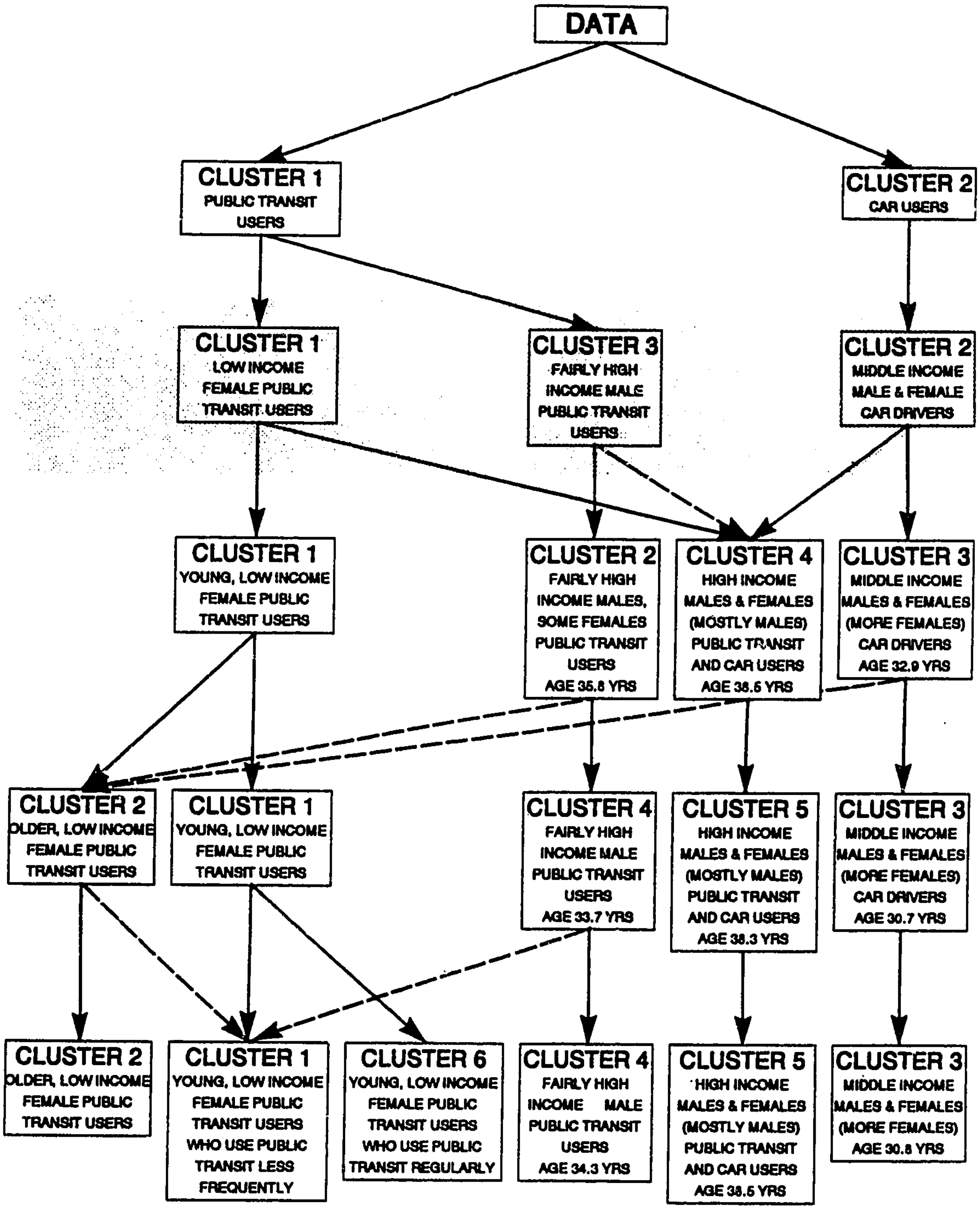

DATA
CLUSTER 1 PUBLIC TRANSIT USERS
CLUSTER 2 CAR USERS
CLUSTER 1 LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 3 FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS
CLUSTER 2 MIDDLE INCOME MALE & FEMALE CAR DRIVERS
CLUSTER 1 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 2 FAIRLY HIGH INCOME MALES, SOME FEMALES PUBLIC TRANSIT USERS AGE 35.8 YRS
CLUSTER 4 HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS AGE 38.6 YRS
CLUSTER 3 MIDDLE INCOME MALES & FEMALES (MORE FEMALES) CAR DRIVERS AGE 32.9 YRS
CLUSTER 2 OLDER, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 1 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 4 FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS AGE 33.7 YRS
CLUSTER 5 HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS AGE 38.3 YRS
CLUSTER 3 MIDDLE INCOME MALES & FEMALES (MORE FEMALES) CAR DRIVERS AGE 30.7 YRS
CLUSTER 2 OLDER, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 1 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS WHO USE PUBLIC TRANSIT LESS FREQUENTLY
CLUSTER 6 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS WHO USE PUBLIC TRANSIT REGULARLY
CLUSTER 4 FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS AGE 34.3 YRS
CLUSTER 5 HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS AGE 38.5 YRS
CLUSTER 3 MIDDLE INCOME MALES & FEMALES (MORE FEMALES) AGE 30.8 YRS

Figure 6.5 Cluster Characteristics from the Three Cluster Solution

income male public transit users are somewhat lower than those for the female public transit users.

Cluster 1, the female public transport users cluster, is the largest cluster. Cluster 3, the higher income, male, public transport users cluster, is the next largest cluster. Based on a sample size of 532 individuals these clusters comprise approximately 76.1% of the sample (50.3% and 25.8%, respectively). Again, because of the large proportion of public transit users in the sample data, these percentages are likely inflated in comparison to the percentages that actually exist in the population of individuals making the home-to-work trip to the Edmonton CBD. Cluster 2 is approximately the same size as cluster 3.

Examination of the membership distribution graphs suggests that all three clusters are "fairly crisp". The Cluster Entropy values indicate that cluster 1 is the least fuzzy cluster, and clusters 2 and 3 have approximately equal "fuzziness".

6.2.3 Four Cluster Results

The four cluster result was derived with the addition of the age and job prestige variables to those variables used to obtain the three cluster solution. Figure 6.6 highlights the characteristics of these clusters.

It is again interesting to note that the four best cluster result contains three clusters (clusters 1,2, and 3) whose characteristics are very similar to clusters developed

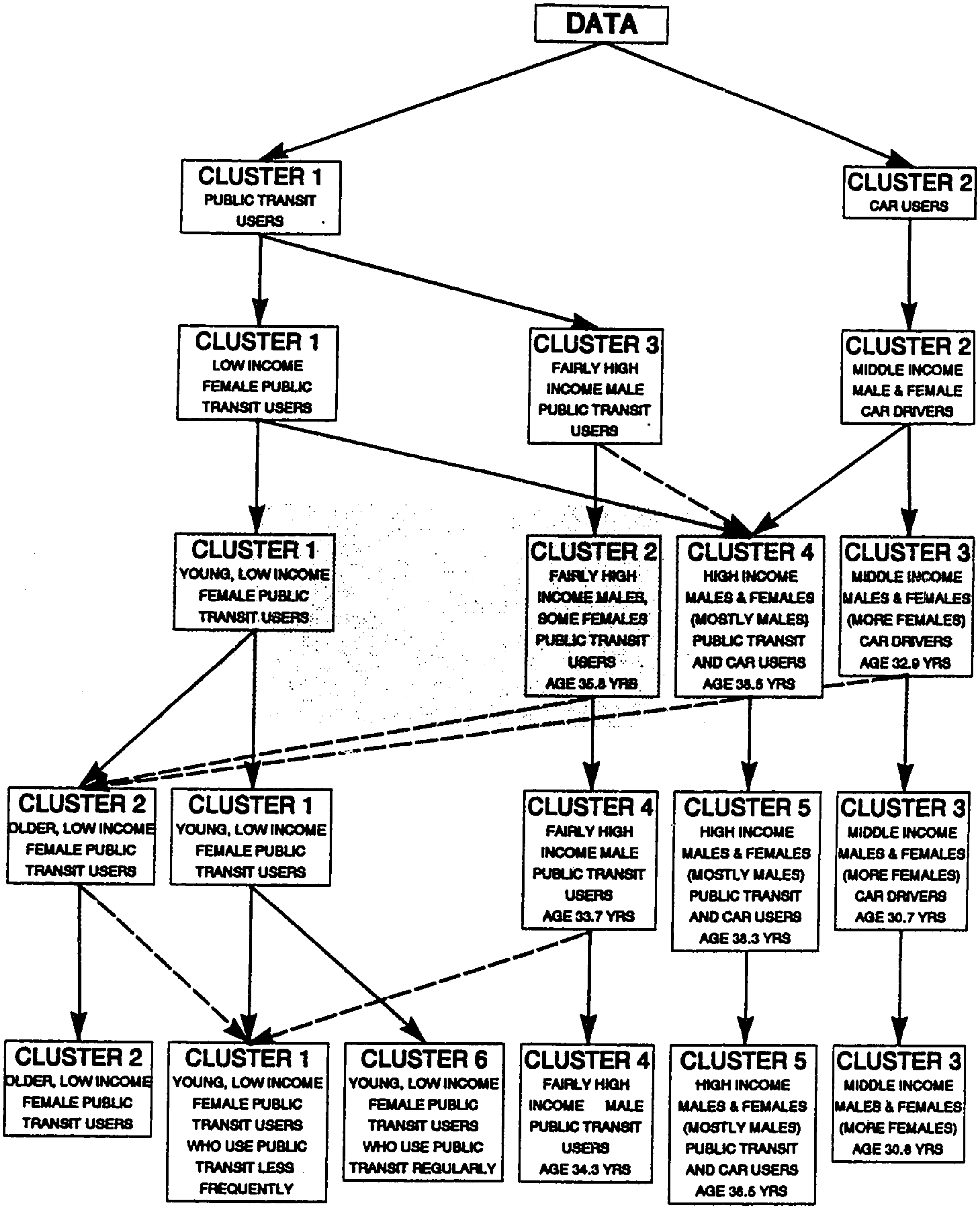

DATA
CLUSTER 1
PUBLIC TRANSIT USERS
CLUSTER 2
CAR USERS
CLUSTER 1
LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 3
FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS
CLUSTER 2
MIDDLE INCOME MALE & FEMALE CAR DRIVERS
CLUSTER 1
YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 2
FAIRLY HIGH INCOME MALES, SOME FEMALES PUBLIC TRANSIT USERS
AGE 35.8 YRS
CLUSTER 4
HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS
AGE 38.5 YRS
CLUSTER 3
MIDDLE INCOME MALES & FEMALES (MORE FEMALES) CAR DRIVERS
AGE 32.9 YRS
CLUSTER 2
OLDER, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 1
YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 4
FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS
AGE 33.7 YRS
CLUSTER 5
HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS
AGE 38.3 YRS
CLUSTER 3
MIDDLE INCOME MALES & FEMALES (MORE FEMALES) CAR DRIVERS
AGE 30.7 YRS
CLUSTER 2
OLDER, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 1
YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS WHO USE PUBLIC TRANSIT LESS FREQUENTLY
CLUSTER 6
YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS WHO USE PUBLIC TRANSIT REGULARLY
CLUSTER 4
FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS
AGE 34.3 YRS
CLUSTER 5
HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS
AGE 38.5 YRS
CLUSTER 3
MIDDLE INCOME MALES & FEMALES (MORE FEMALES)
AGE 30.8 YRS

Figure 6.6 Cluster Characteristics from the Four Cluster Solution

in the three cluster solution. The main difference between these clusters and clusters 1,3, and 2, in the three cluster solution, is that in the four cluster solution the income variable value has, for each, decreased. The other cluster properties have remained more-or-less the same, and therefore will not be re-described here.

For the most part it seems that the higher income individuals who were removed from the previous three clusters now make up the newly defined cluster 4. Cluster 4 contains relatively high income (average income approximately 33,450 $/yr), males and females (slightly more males), whose usual mode choice is split almost equally between public transit and automobile. The average age for this group is approximately 38.5 yrs. It is interesting to note the relatively low travel time estimations for this cluster; the values are similar to those of the female public transit users cluster. This is an interesting and unexpected result considering the socioeconomic and mode choice characteristics of cluster four.

The addition of the job prestige variable basically duplicated the results obtained with the income variable. As was expected, the highest prestige jobs (jobs associated with lower JOB variable values) had the highest average incomes and as the average income decreased so too did the prestige associated with the job.

Both the Cluster Entropy measure and the membership graphs indicate that the cluster memberships for these clusters are fuzzier thän the memberships for the previously analysed clusters. This trend will be discussed in more detail in Section 6.3.2. Here it will suffice to say that cluster 1, the female public transport users cluster, is the least fuzziest cluster. The other three clusters have approximately the same fuzziness as indicated by their Cluster Entropy values.

6.2.4 Five Cluster Results

The five cluster solution was also considered a "best" solution for the 10 dimension case. The characteristics of these clusters are highlighted in Figure 6.7.

Again, as has been the trend in the cluster splitting process, the five best clusters contain four previously defined clusters plus one new one. It seems that the newly created cluster has largely been derived from splitting the young, low income female public transit users cluster; cluster 1, in the four cluster solution. Basically, this cluster has divided into (1) a cluster of young (age approximately 24 yrs.), low income, female public transit users and (2) a cluster of older (age approximately 44.5 yrs), low income, female public transit users. There are perceptual differences between these clusters. The older female public transport users' estimations of both their walk and wait times are relatively high. The estimations are in the same range as the car users of cluster 3. It is surprising that this large group of public transit users

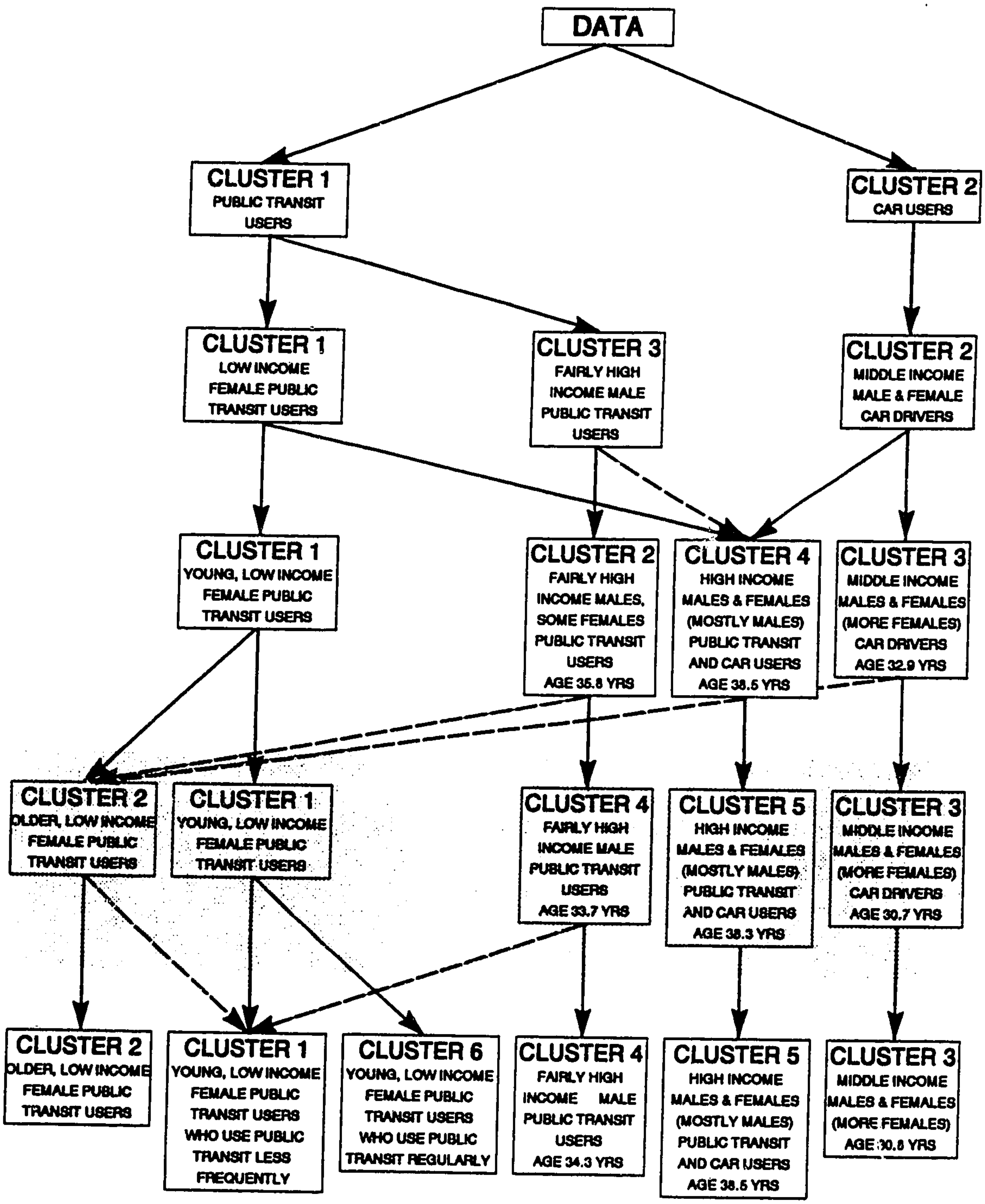

DATA
CLUSTER 1 PUBLIC TRANSIT USERS
CLUSTER 2 CAR USERS
CLUSTER 1 LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 3 FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS
CLUSTER 2 MIDDLE INCOME MALE & FEMALE CAR DRIVERS
CLUSTER 1 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 2 FAIRLY HIGH INCOME MALES, SOME FEMALES PUBLIC TRANSIT USERS AGE 35.8 YRS
CLUSTER 4 HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS AGE 38.5 YRS
CLUSTER 3 MIDDLE INCOME MALES & FEMALES (MORE FEMALES) CAR DRIVERS AGE 32.9 YRS
CLUSTER 2 OLDER, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 1 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 4 FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS AGE 33.7 YRS
CLUSTER 5 HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS AGE 38.3 YRS
CLUSTER 3 MIDDLE INCOME MALES & FEMALES (MORE FEMALES) CAR DRIVERS AGE 30.7 YRS
CLUSTER 2 OLDER, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 1 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS WHO USE PUBLIC TRANSIT LESS FREQUENTLY
CLUSTER 6 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS WHO USE PUBLIC TRANSIT REGULARLY
CLUSTER 4 FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS AGE 34.3 YRS
CLUSTER 5 HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS AGE 38.5 YRS
CLUSTER 3 MIDDLE INCOME MALES & FEMALES (MORE FEMALES) AGE 30.8 YRS

Figure 6.7 Cluster Characteristics from the Five Cluster Solution

would perceive similar values for their public transit travel times as do individuals who usually drive, especially considering, that these automobile drivers have had the highest perceptions throughout this analysis. The perceptions for cluster 1, the young female public transit users are much lower; they are in the same range as those of the young female public transit users in the four cluster solution.

The other three cluster defined here, clusters 3,4, and 5 have characteristics similar to clusters 3,2, and 4 from the four cluster solution. For all three clusters, however, the GENDER value has decreased slightly (indicating a larger proportion of males) and the INCOME value has increased slightly. Likely the females who had earlier belonged to these "middle-aged" clusters have been re-dispersed to the newly created cluster 2.

6.2.5 Six Cluster Results

The addition of the mode frequency variable resulted in the creation of one more cluster. This cluster, cluster 1 in the six cluster solution, contains low income, fairly young (approximately 26.5 yrs) females who, on average, use public transport less often than do members of the other clusters. These individuals likely belonged to clusters 1,2, and 4 in the previous five cluster solution. For further cluster characteristic details see Figure 6.8.

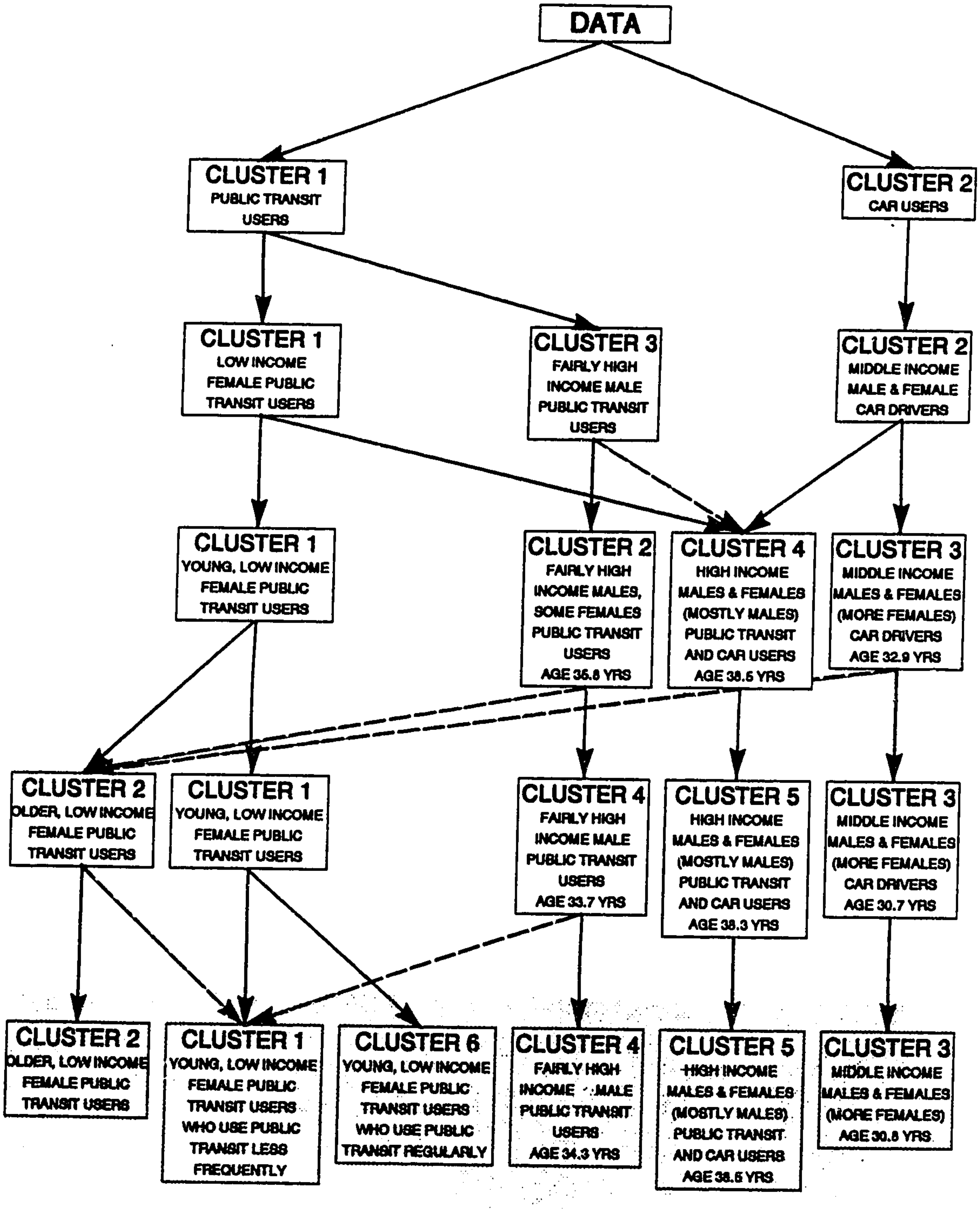

DATA
CLUSTER 1 PUBLIC TRANSIT USERS
CLUSTER 2 CAR USERS
CLUSTER 1 LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 3 FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS
CLUSTER 2 MIDDLE INCOME MALE & FEMALE CAR DRIVERS
CLUSTER 1 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 2 FAIRLY HIGH INCOME MALES, SOME FEMALES PUBLIC TRANSIT USERS AGE 35.8 YRS
CLUSTER 4 HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS AGE 38.5 YRS
CLUSTER 3 MIDDLE INCOME MALES & FEMALES (MORE FEMALES) CAR DRIVERS AGE 32.9 YRS
CLUSTER 2 OLDER, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 1 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 4 FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS AGE 33.7 YRS
CLUSTER 5 HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS AGE 38.3 YRS
CLUSTER 3 MIDDLE INCOME MALES & FEMALES (MORE FEMALES) CAR DRIVERS AGE 30.7 YRS
CLUSTER 2 OLDER, LOW INCOME FEMALE PUBLIC TRANSIT USERS
CLUSTER 1 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS WHO USE PUBLIC TRANSIT LESS FREQUENTLY
CLUSTER 6 YOUNG, LOW INCOME FEMALE PUBLIC TRANSIT USERS WHO USE PUBLIC TRANSIT REGULARLY
CLUSTER 4 FAIRLY HIGH INCOME MALE PUBLIC TRANSIT USERS AGE 34.3 YRS
CLUSTER 5 HIGH INCOME MALES & FEMALES (MOSTLY MALES) PUBLIC TRANSIT AND CAR USERS AGE 38.5 YRS
CLUSTER 3 MIDDLE INCOME MALES & FEMALES (MORE FEMALES) AGE 30.8 YRS

Figure 6.8 Cluster Characteristics from the Six Cluster Solution

With the exception of cluster 1, the MODEFREQ values for the other clusters were almost the same. As was discussed in Section 4.4.3 this was likely due to the fact that the data, on which this research was based, contained "regular" mode users only. It is unfortunate that the database did not contain individuals with more mode usage variation, as the results with the "regular" users were as one would predict. These results being that, except for cluster 2 (the older, female, public transit users), those individuals from cluster 1 who used public transit less often had higher public transit time perceptions than did those individuals from clusters 4 and 6, the transit users who used public transit more frequently.

The cluster characteristics for the six clusters are summarized below:

- Cluster 1: is characterized by relatively low income (avg. income 18,300 $/yr), fairly young (avg. age 26.7 yrs), female public transit users who, on average, use public transit less often than do the individuals who belong to the other public transit using clusters.
- Cluster 2: is characterized by low income (avg. 15,900 $/yr), older (avg. 45.8 yrs) female public transit users.
- Cluster 3: is characterized by middle-aged (avg. 30.8 yrs) male and female auto users, who earn, on average, 21,800 $/yr.

- Cluster 4: is characterized by middle-aged (avg. 34.3 yrs) male public transit users, who earn, on average 26,650 $/yr.
- Cluster 5: is characterized by slightly older (avg. 38.5 yrs), males and females [illegible] males), who have, on average, the highest [illegible] of the sample data (avg. income 36,200 $/yr). Mode usage for this cluster is almost split evenly between public transit and private vehicle. This is the smallest cluster (approximately 10.0% of the sample).
- Cluster 6: is characterized by low income (avg. 15,950 $/yr), young (avg. 24.0 yrs), female public transit users, who use public transit more often than do those females of cluster 1. This is the largest (27.5% of the sample) and least fuzzy cluster.

The perceptions of the six clusters have been compared on the line graphs in Figures 6.9 and 6.10. According to the perception ratio and perception difference measures, clusters 4, 5, and 6 have the lowest public transit walk and wait time estimates. For clusters 4 and 6 this is a logical result as members of these clusters are regular public transit users and, therefore, one would expect their public transit walk and wait time estimates to be accurate. Somewhat surprising, is the accurate transit time estimations for the individuals of cluster 5, as only approximately one-half of this high income cluster use

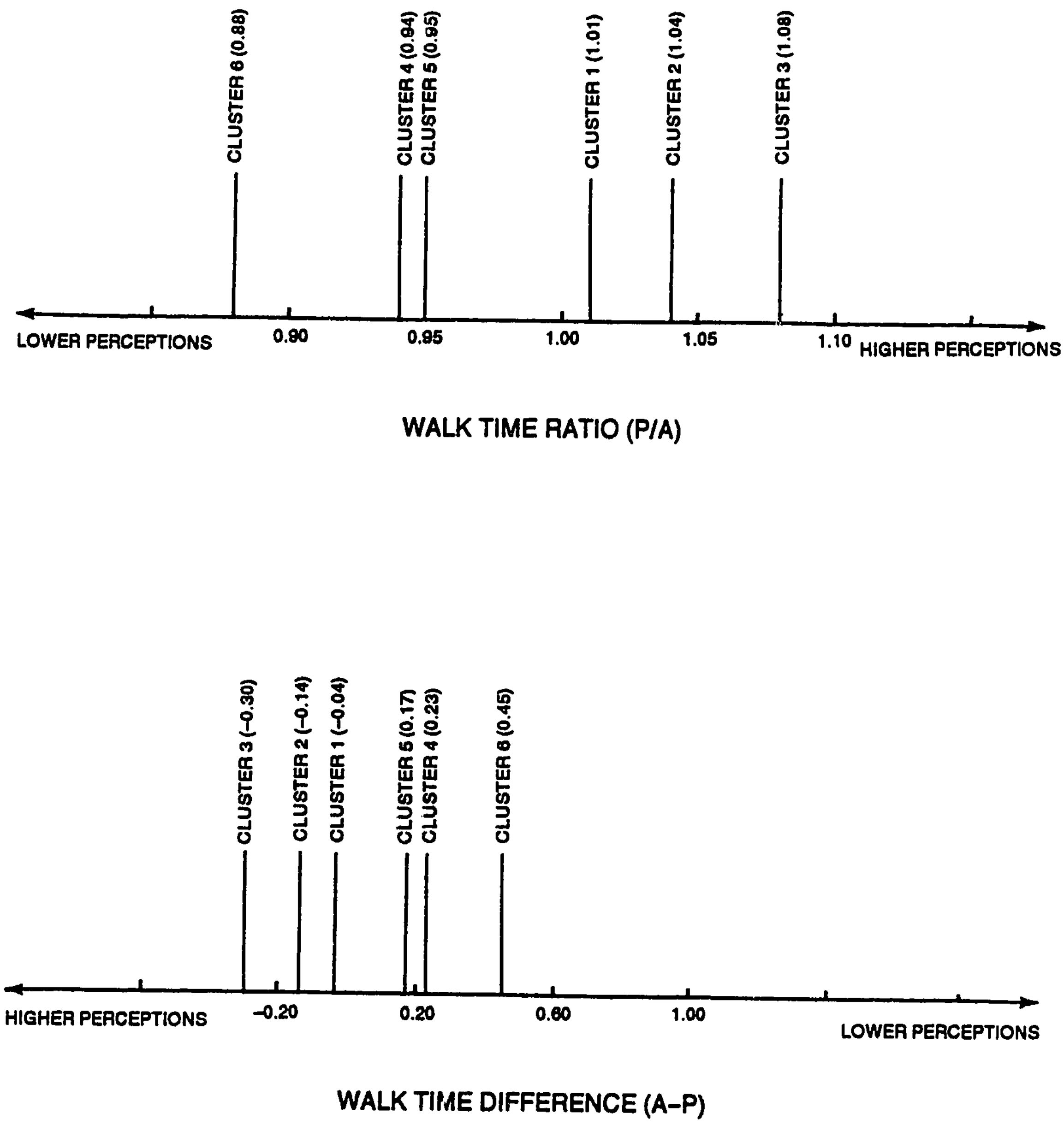


Figure 6.9 Comparison of Walk Time Ratios and Walk Time Differences for the Six Cluster Solution (Based on Average Cluster Characteristics)

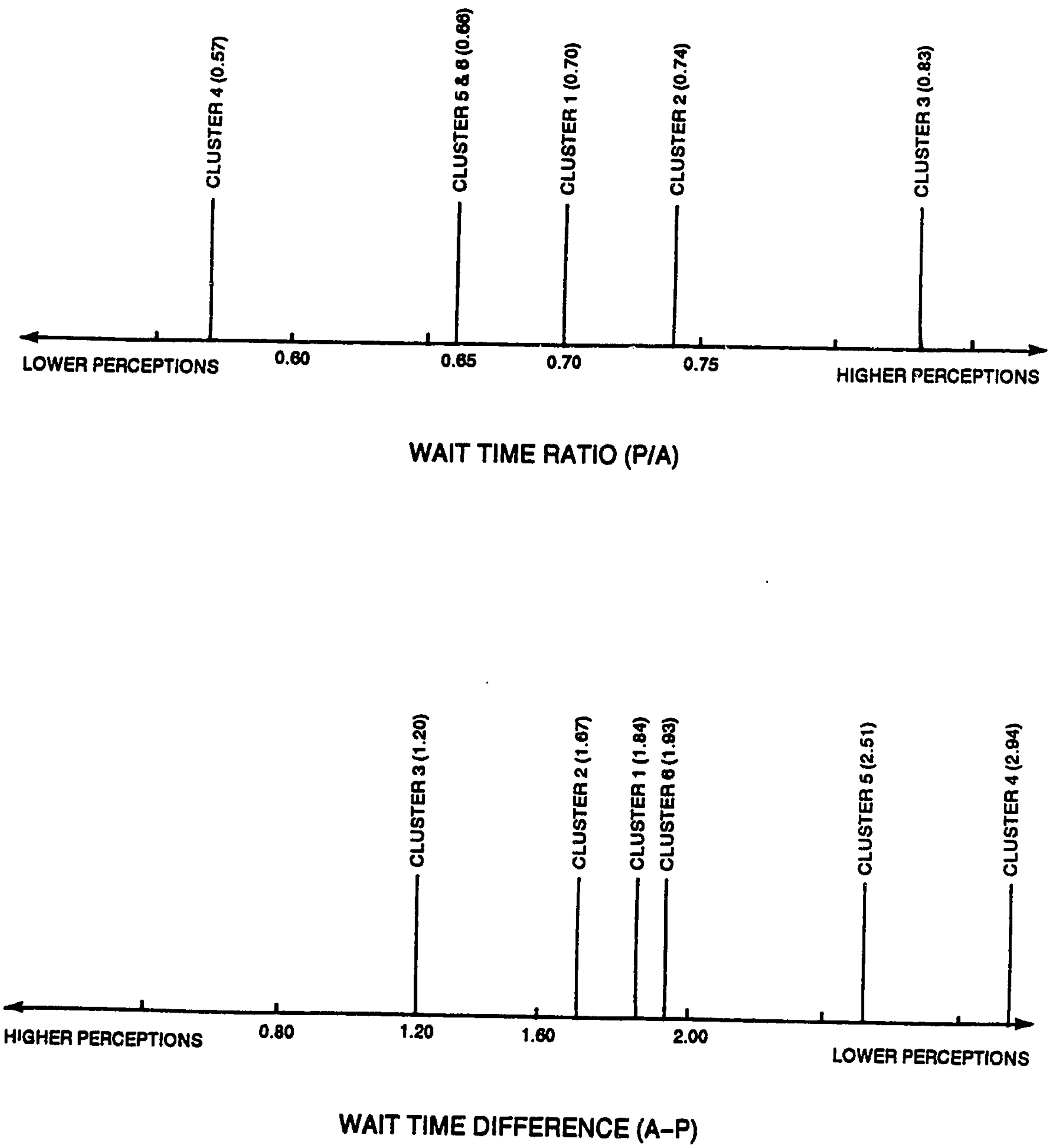


Figure 6.10 Comparison of Wait Time Ratios and Wait Time Differences for the Six Cluster Solution (Based on Average Cluster Characteristics)

public transit regularly.

Conversely, the individuals in clusters 1, 2, and 3 have the highest walk time and wait time estimates. Logically, the absolute highest time estimations are associated with the private vehicle users of cluster 3. This suggests a lack of knowledge of public transit service attributes by these individuals. These results coincide with results obtained in past research (Quarmby, 1967; O'Farrell and Markham, 1974; Heggie, 1976; Meyburg and Brog, 1981). Higher public transit wait time and walk time estimates might also be expected from individuals in cluster 1, as they are, on average, less frequent users of public transit. This same result, however, was unexpected for the regular transit users of cluster 2. It is surprising that this group of older, low income, females would consistently have such high estimates for their public transit walk and wait time perceptions. The use of the constant walking speed might explain the bias for the walk time overestimations. Cluster 2 contains, on average, older members who likely have slower average walking speeds relative to the other clusters. The use of a constant walking speed for all clusters likely results in a walk time value for this cluster that is too fast relative to the other cluster's walk time estimations. This, in turn, results in an increased difference between the perceived and actual walk time values. No explanation is hypothesized for the high wait time estimations.

As was discussed earlier, because no standard deviation measure exists for cluster centre values, differences between the average cluster perceptions, as measured by the perception ratio and difference values, cannot be statistically tested.

6.3 Observations

6.3.1 Cluster Splitting Trends

It is observed that as the number of dimensions was increased so too did the "best" number of clusters, as specified by the Partition Entropy measure, increase. This result may be an indication that a particular number of well-separated clusters do not exist. Rather what likely exist are not-so-well-separated clusters that are defined only as more information is made available for the clustering procedure. The presence of not-so-well-separated clusters does not indicate a less valid solution, if this is a property of the data. In fact, the presence of not-so-well-separated clusters is justification for the use of the Fuzzy Cluster method.

From Figure 6.2 it is observed that new clusters were created by combining and splitting the fewer number of clusters of the previous solutions. Intuitively, it seems logical that the presence of previously defined clusters in the "new" solutions suggests stability for these particular clusters in the data. However, to the best of the author's knowledge there exists no measure for cluster stability. Instead, one is left to judge the reasonableness of a solution by intuition only. All cluster solutions described in the previous section are judged to be reasonable.

6.3.2 Cluster Fuzziness

Both the Cluster Entropy measure and the membership distribution graphs indicate that as the "best" number of clusters describing the data structure increases, so too does the fuzziness of the clusters. Two possible explanations for why this might occur are:

1. As was discussed in Chapter 5, the Partition Entropy of a solution tends to increase as the number of clusters increases (Windham, 1981). Since the Partition Entropy measure and the Cluster Entropy measure are both a function of the entropy values, which are a function of the object's membership, one would expect to see an increase in the cluster's entropy as the Partition Entropy increases.
2. In addition to this, it was also noted in Chapter 5 that as the number of dimensions increases so too do the Partition Entropy values associated with that solution. Again, with this increase in the Partition Entropy one would expect to see an increase in the Cluster Entropy values.

The solutions, then, associated with the higher cluster numbers are not considered to be less valid solutions because of their increased fuzziness.

6.3.3 Assumptions for Actual Walk and Wait Time Calculations

The following observations are made regarding the actual walk and wait time assumptions used in this research:

1. Examination of the differences between the actual walk times and the perceived walk times suggest that the assumed walk velocity was a reasonable one. In the six cluster solution the cluster with the largest average walk time overestimation was cluster 3 with an average overestimation of eighteen seconds. The cluster with the largest average walk time underestimation was cluster 6 with an underestimation of twenty-seven seconds. The difference between these two extremes is only forty-five seconds.
2. Examination of the differences between the actual wait times and the perceived wait times raises some doubt as to the validity of the assumed wait time model. It is highly unlikely that most individuals underestimated their public transit wait time as is suggested by these results. More likely, the underlying assumption that people are arriving randomly at the public transit stop is incorrect. Assuming that people arrive in a more systematic fashion would result in a shorter, and seemingly more accurate, predicted wait time values. Since,

however, the predicted wait time values were used for comparing the clusters' perceptions only, the results associated with these values are considered to be valid.

6.4 Further Analysis of the Six Cluster Solution Using the Concept of Cluster Cores

The next section examines the characteristics of the "cluster cores". The six cluster solution was used for this analysis because: (1) this solution seems to have evolved from the smaller cluster solutions and (2) all eleven variables are represented in this solution.

6.4.1 Cluster Core Definition

Bezdek (1981) defines the "core" of a cluster as the "crisp" set of objects which belong to a cluster when a threshold membership value of γ is specified. This concept is illustrated in Figure 6.11. In this figure the shaded area represents the cluster core for a threshold value of 0.9. All objects which have membership values greater than or equal to 0.9 belong to this cluster core. In the same manner cluster cores could be defined for threshold values of 0.8, 0.7, etc.

6.4.2 Cluster Core Analysis

In this research, cluster cores were developed for threshold values of 0.8, 0.7, 0.6, and 0.5. From visual

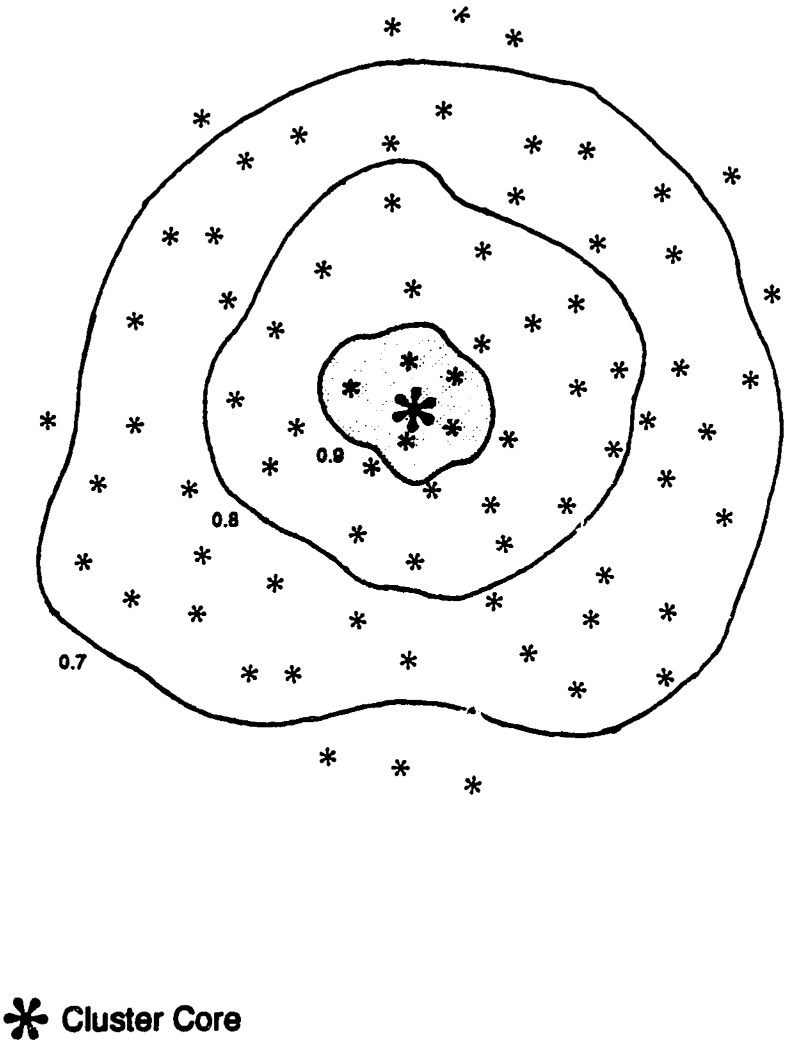


Figure 6.11 Concept of Cluster Cores

examination of the cluster membership values it was determined that, for the six cluster solution, a 0.5 membership value was still a reasonably high membership value and that results obtained using this value would accurately represent the cluster characteristics.

6.4.3 Core Size Elasticity

First the effect of changing γ on the core sizes was investigated. The histogram in Figure 6.12 shows the number of objects which belong to each of the cluster cores for each of the threshold values. The percentages illustrate what proportion of the 0.5 cluster core each threshold range occupies. For example, in the 0.5 cluster core for cluster 1, 52.4% of the objects have memberships greater than 0.8, 23.0% have memberships between 0.7 and 0.8, 16.4% have memberships between 0.6 and 0.7, and 8.2% have memberships between 0.5 and 0.6. This information essentially quantifies that obtained from the membership distribution graphs, for the threshold ranges investigated.

This graph may also be used to make inferences regarding cluster size. Based on the previous definition of cluster size, that larger clusters contain a larger number of high membership objects, cluster 6 would clearly be labelled the largest cluster and cluster 5 the smallest cluster. These same results were obtained using the cluster size measure developed in this research.

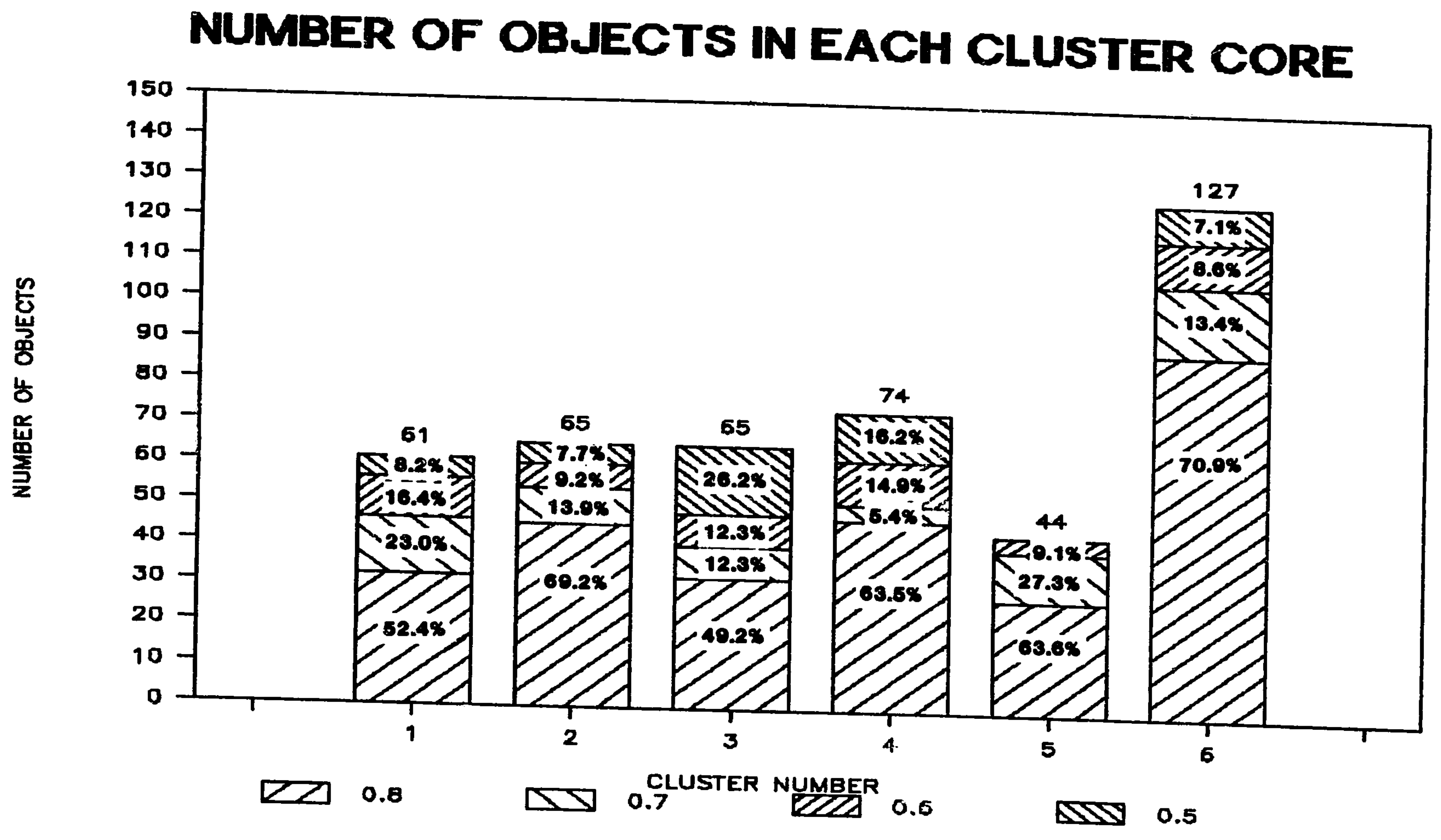


Figure 6.12 Number of Objects in Each Cluster Core for 0.5, 0.6, 0.7 and 0.8 Threshold Values

The line graph in Figure 6.13 illustrates the responsiveness, or elasticity, of the cluster core size to a change in threshold value. For clusters 1,2,4 and 6, the graph indicates a similar core size increase for the same threshold value decrease. The disproportionate increase in the cluster core size for cluster 3, between the 0.6 and 0.5 threshold values, indicates the fuzzier nature of this cluster (this was the fuzziest cluster according to the Cluster Entropy measure). Oppositely, the crisp nature of cluster 5 is illustrated by the presence of no objects in this region (cluster 5 was one of the least fuzzy clusters according to the Cluster Entropy measure).

Information such as this may be very important for market strategy development. One may argue that individuals who do not clearly belong to a particular cluster may be more susceptible to marketing strategies designed to change the characteristics typically exemplified by the cluster. This argument suggests that for better marketing investment returns, marketing strategies should be aimed at fuzzier clusters, or in this context fuzzier market segments. In this research cluster 3 may be an ideal candidate. This concept is discussed in further detail in Chapter 8.

6.4.4 Core Characteristics

Characteristics of the cluster cores were examined by plotting the distribution of the characteristic values for the same varying values of γ. The distribution graphs for

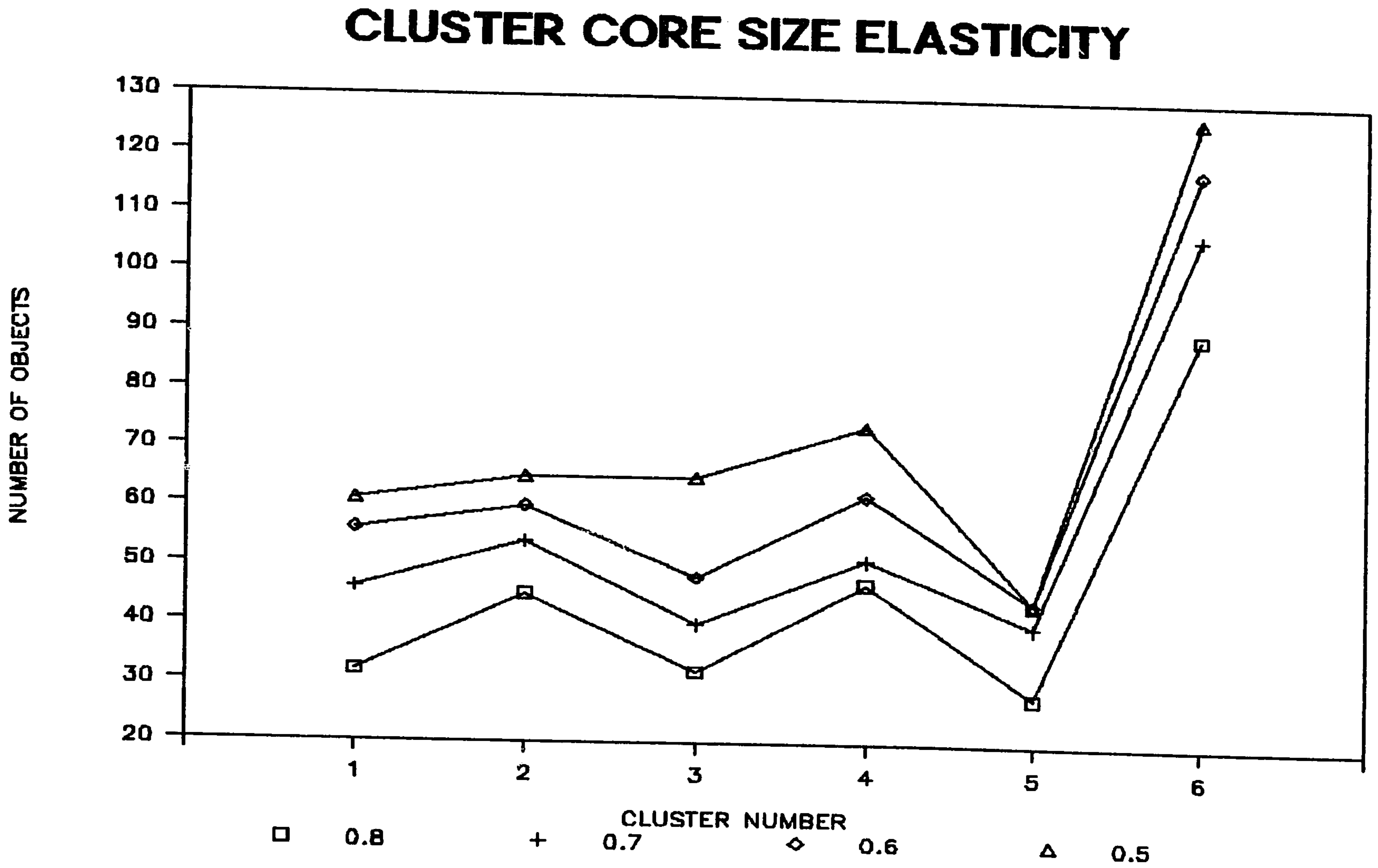


Figure 6.13 Cluster Core Size Elasticity

The socioeconomic and mode choice variables, for all six clusters, are contained in Appendix C. These graphs for two typical clusters, clusters 2 and 5, have been included in this chapter as Figures 6.14 and 6.15.

What is first evident from examination of these graphs is the homogeneity of the cluster characteristics. One would definitely conclude that cluster 2 contains relatively low income, female, public transit users; whereas, cluster 5 contains both males and females (mostly males) who earn higher incomes, and who use, almost equally, both private vehicles and public transit for their morning home-to-work trip. These characteristics for the other clusters are similarly evident from their characteristic distribution graphs.

Distribution histograms were also produced for the walk and wait time ratios and the walk and wait time difference measures. All graphs are contained in Appendix D. The walk and wait time difference plots for clusters 2 and 5 have been included in this chapter as Figure 6.16.

Examination of these "perception" graphs indicates that the perceptual differences between the clusters are not as clearly defined as are the socioeconomic and mode choice characteristic differences. From visual examination of the walk and wait time differences graphs for clusters 2 and 5, it is evident that the older female public transit users of cluster 2 perceive their public transit wait time somewhat longer, on average, than do the individuals of cluster 5.

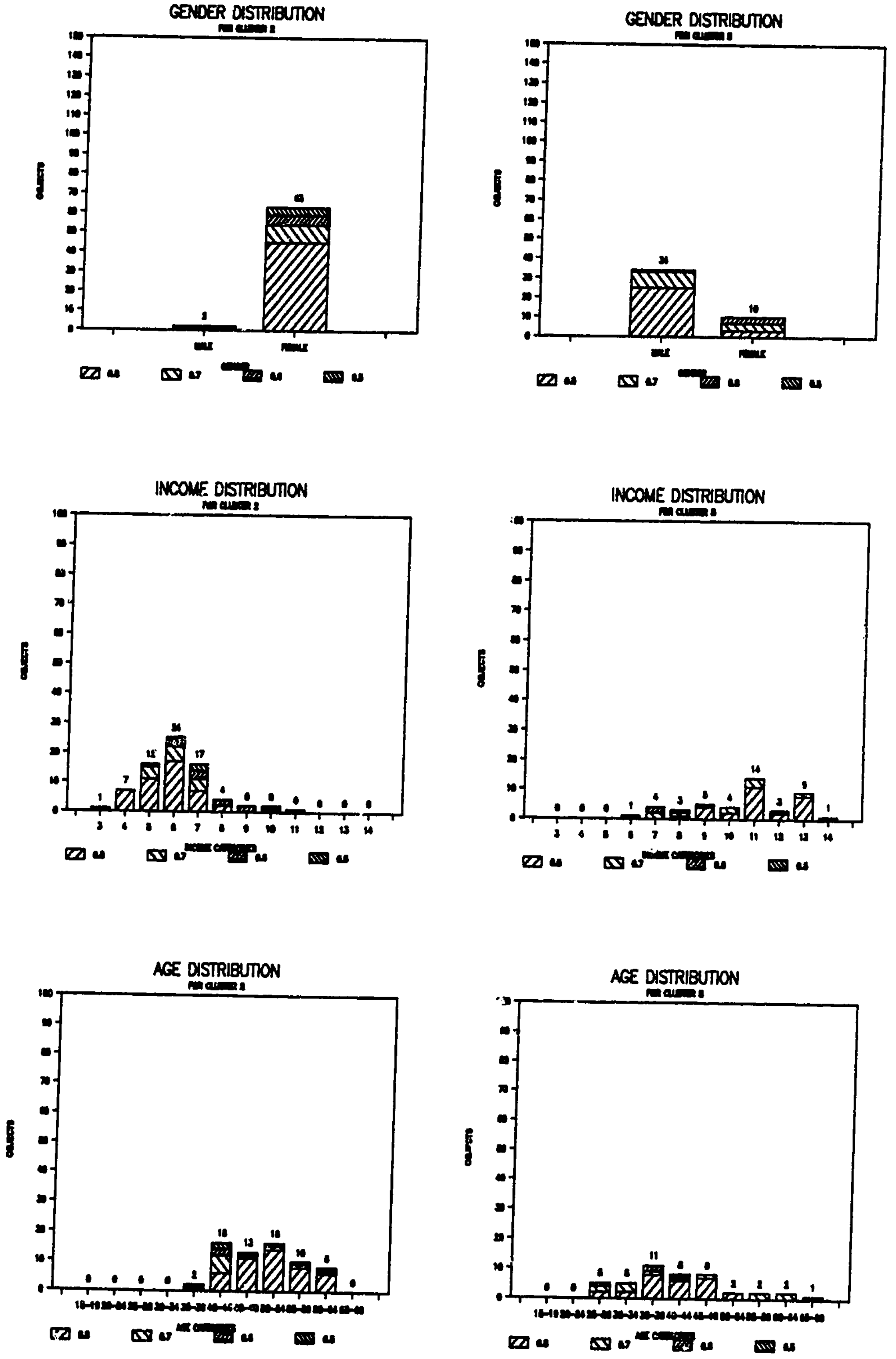


Figure 6.14 Gender, Income, and Age Distributions for Cluster Cores 2 and 5

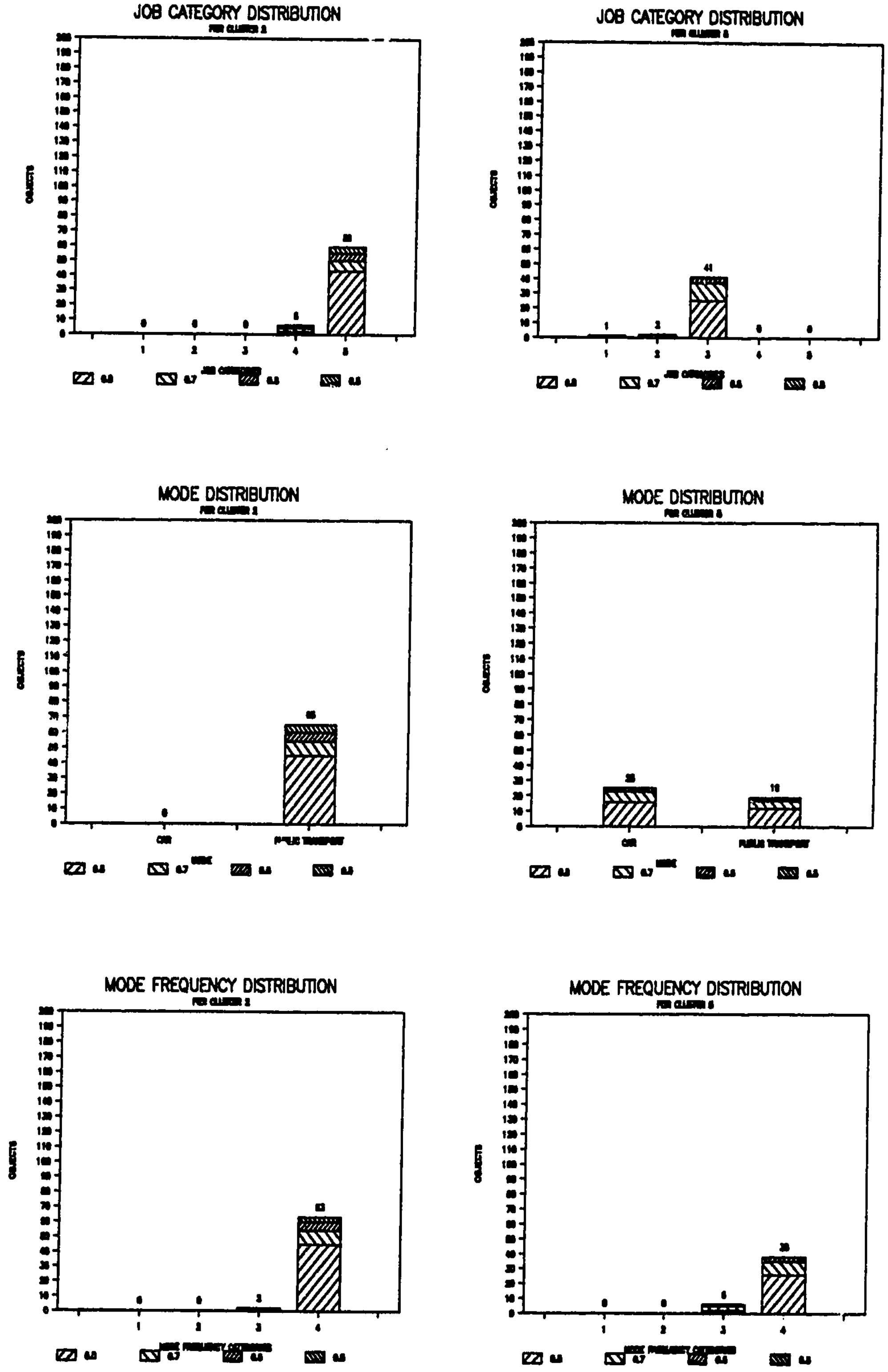


Figure 6.15 Job, Mode, and Mode Frequency Distributions for Cluster Cores 2 and 5

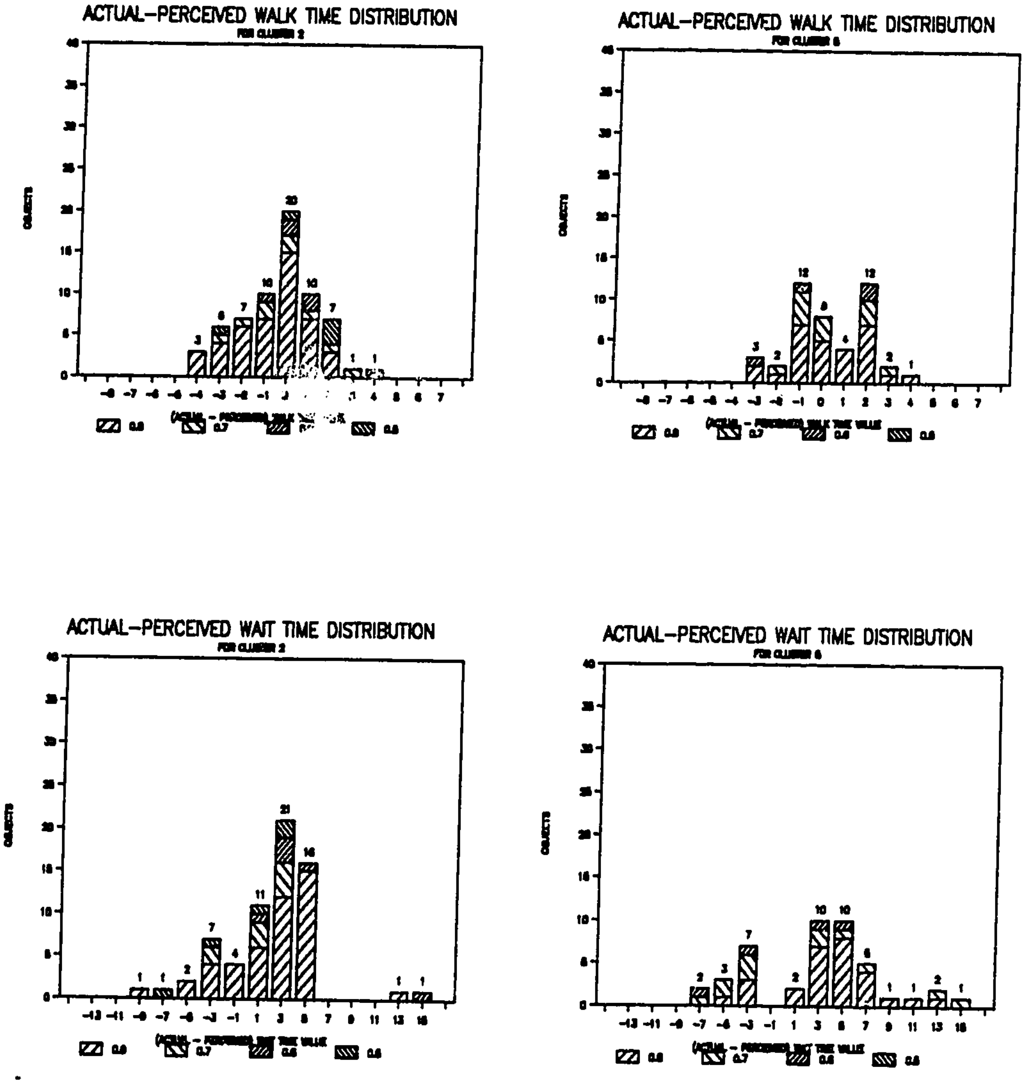


Figure 6.16 Walk Time and Wait Time "Perception Difference" Distributions for Cluster Cores 2 and 5

With regard to walk times, again those individuals in cluster 2 perceive their walk times as being longer, however, the average difference here is much smaller.

An advantage of core analysis is that now a "crisp" membership is defined and, as a result, the dispersion properties of the cluster characteristics can be examined. These dispersion properties, particularly standard deviation values, can be used to test whether differences between cluster characteristics are significantly different.

Because the socioeconomic and mode choice differences between clusters were evident from visual examination of the characteristic graphs, statistical differences between these variables were not tested. Differences between cluster perception measures were, however, tested. The statistical test used to test whether the cluster's mean perceptions were significantly different was the student's t-test. The mean and standard deviation values required for this test were taken for the 0.8 threshold results. The 0.8 threshold mean and standard deviation values are contained in Table 6.2.

Use of the student's t-test assumes that the variables for which differences are being tested are normally distributed with a constant variance (Box, Hunter, and Hunter, 1978). Normal probability plots were produced to test the normality of the walk and wait time ratio and difference values for each cluster. These plots were produced using the MIDAS program on the University of

CLUSTER	DEGREES OF FREEDOM	MEAN VALUES				STANDARD DEVIATIONS			
		WALK RATIO	WAIT RATIO	ABS. WALK	ABS. WAIT	WALK RATIO	WAIT RATIO	ABS. WALK	ABS. WAIT
CLUSTER 1	31	1.4291	0.8933	-0.2691	1.5781	1.0044	0.7231	1.5350	2.7711
CLUTSER 2	44	1.4889	0.8401	-0.6157	1.7444	1.1667	0.6612	1.5577	3.7121
CLUSTER 3	31	1.1909	0.9531	-0.0352	1.7031	0.8803	0.8773	1.4051	3.8560
CLUSTER 4	46	1.0268	0.6905	0.3892	2.7766	0.5030	0.7223	1.4055	3.4981
CLUSTER 5	27	1.1128	0.5966	0.3175	3.5357	0.6712	0.5468	1.8131	4.0687
CLUSTER 6	89	1.0521	0.8126	0.5173	2.1667	0.8878	0.7668	1.6262	3.0731

Table 6.2 Perception Measure Means and Strandard Deviation Values (Core Threshold = 0.8)

Alberta mainframe computer. It is assumed that if the transformed variable values plot on a straight line, the distribution is approximately normal. The normal probability plots for the four perception measures, for the six clusters, are contained in Appendix F. Examination of these graphs suggests: (1) the normality assumption for the walk and wait time ratio distributions is, perhaps, questionable, and (2) the walk and wait time difference distributions are more-or-less normally distributed. Because the t-test remains valid for distributions which deviate moderately from normal, it was still used to test whether perceptual differences between clusters were significant (Geary, 1936). The distributions' deviations from normality were taken into account in the analysis of the t-test results.

The F-test was used to test whether variable variances between cluster pairs were statistically equivalent. The F-test also assumes normality for the distributions whose variances are being compared. This test is more sensitive to departures from normality (Statistical Research Laboratory, 1976). Again, the distributions' deviations from normality were considered during the analysis of the t-test results. F-statistics were calculated for each possible pair of clusters, for the perception difference and ratio distributions. These statistics are contained in Appendix F. Based on the F-statistic results, all variance pairs, except for four of the walk ratio variance comparisons, were found to be statistically equivalent for $\alpha=10\%$.

Other assumptions required for the use of the student's t-test are (Box, Hunter and Hunter, 1978):

1. The values for which the means are being tested have been obtained through random, independent observations.
2. All observations were obtained under similar conditions.

These assumptions are basically true for this analysis. For further information regarding this aspect of the research see Chapter 3.

T-statistics were generated for all pairs of clusters for the perception ratio and difference values. These statistics are contained in Appendix F.

The line graphs in Figures 6.17 and 6.18 rank the mean perceptual measures for each cluster core and indicate which are significantly different at a 90% level. Clusters which are significantly different at this level of significance are connected by arrows.

On the basis of these graphs it was concluded that the t-test results were reasonable. The perceptual differences between clusters at opposite ends of the scales were found to be statistically significant. The extent of these differences were not evident from visual examination of the perception variable distribution plots. Perceptual differences between "more" perceptually similar clusters were not significant.

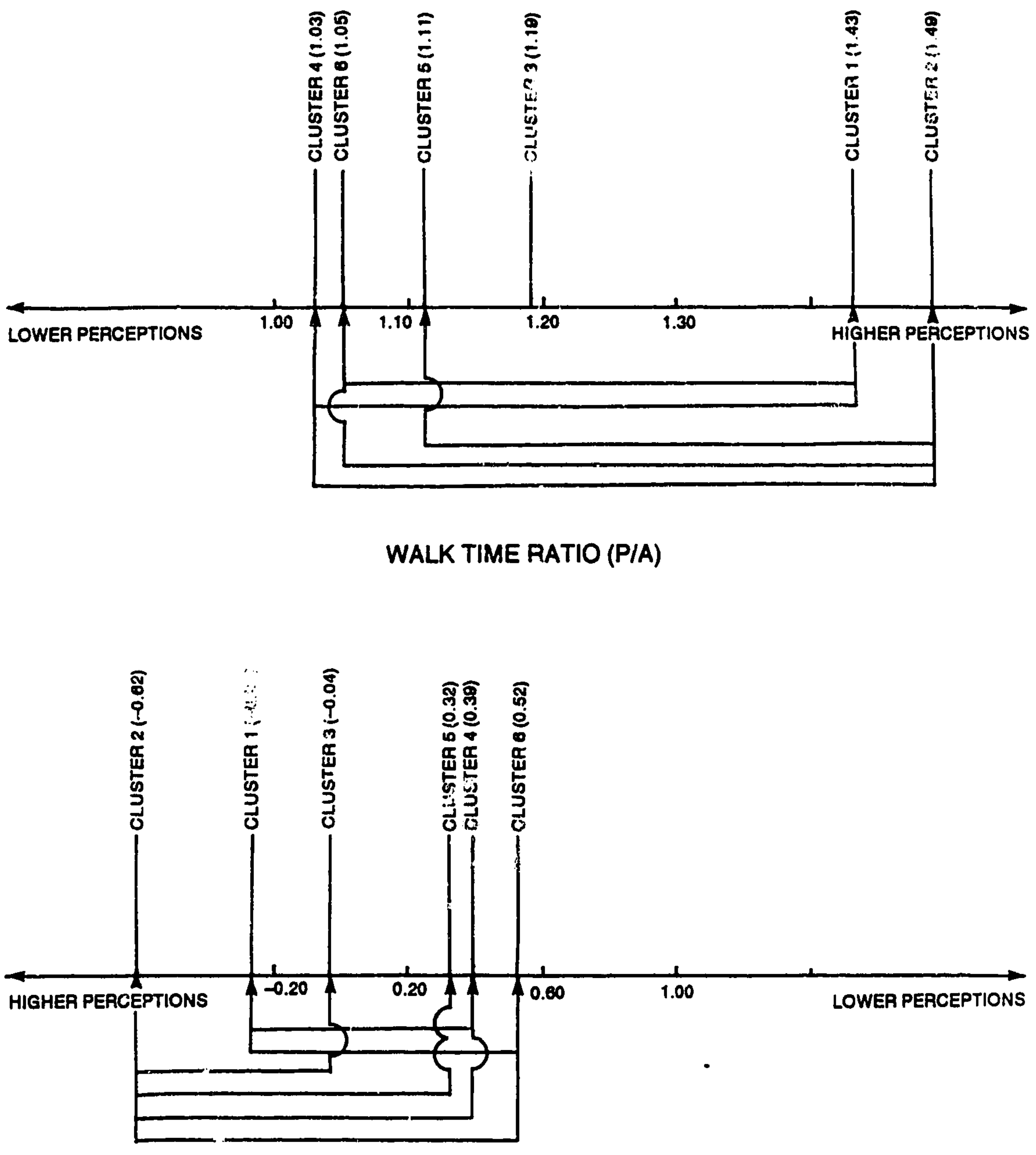


Figure 6.17 Comparison of Walk Time Ratios and Walk Time Differences for the Six Cluster Solution (Based on Cluster Core Results)

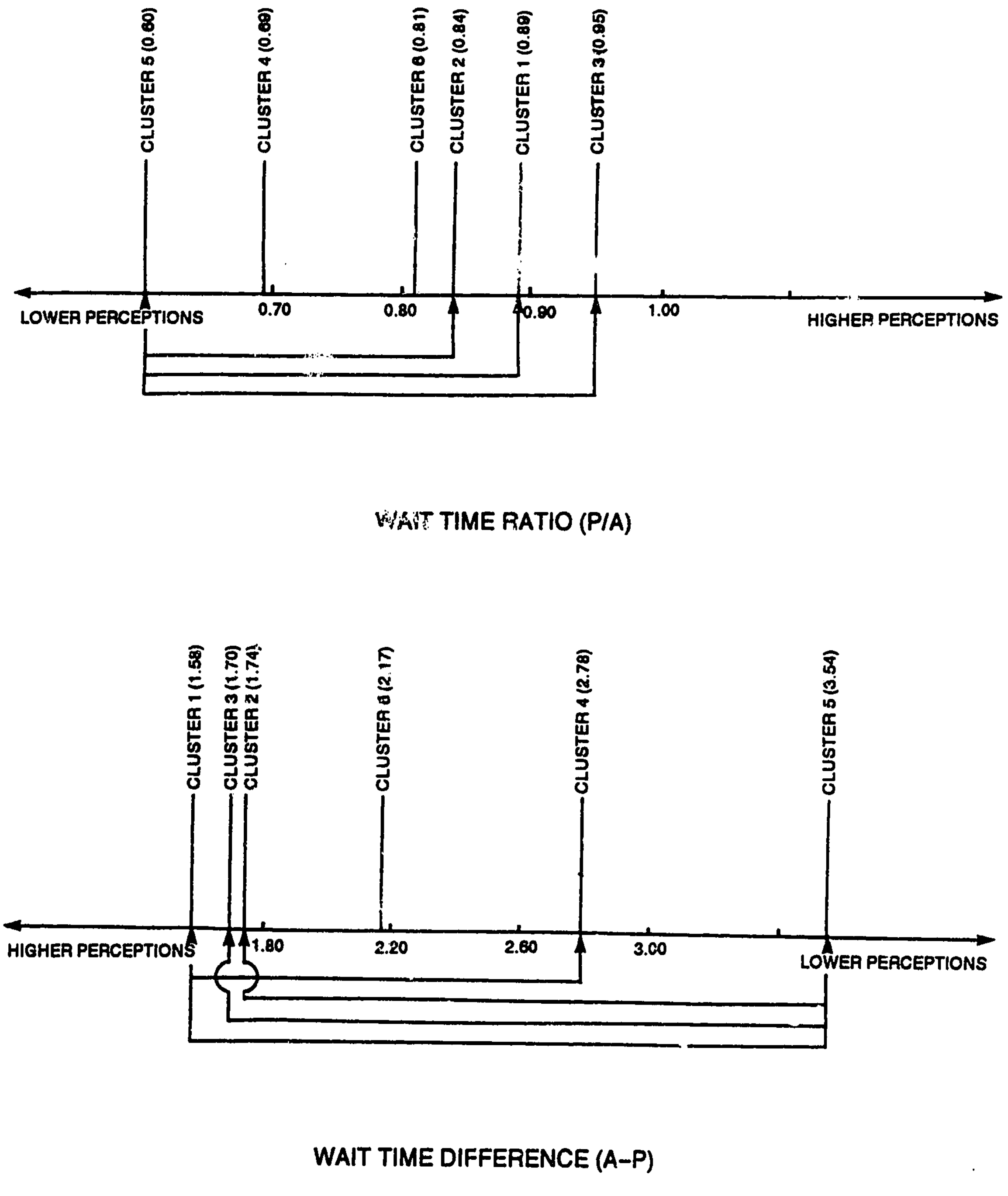


Figure 6.18 Comparison of Wait Time Ratios and Wait Time Differences for the Six Cluster Solution (Based on Cluster Core Results)

With regard to the ranking of the clusters' perceptions, it was observed that the results based on the cluster core analysis were similar to those obtained from the analysis using the average cluster characteristic values. For the description of these results see Section 6.2.5.

7. FURTHER ANALYSIS OF TRAVELERS' PECEPTIONS USING LINEAR REGRESSION

Because of the relative newness of the Fuzzy Cluster Method, further examination of travelers' perceptions of public transit travel times and the variables thought to influence these perceptions was undertaken using a more "traditional" approach. Since linear regression is a method of analysis which might, traditionally, have been used for investigation of this this type of problem it was employed.

This chapter [illegible]ts the development of two multivariate lin[illegible]ression models; one model to predict travelers' perceived walk times and one to predict travelers' perceived wait times. Following this documentation is a brief discussion comparing the linear regression results with those obtained from the Fuzzy Cluster analysis.

7.1 Perceptions Theory

Travelers' perceptions of travel times are likely a function of many variables. Some of these are examined in the following discussion.

Past research suggests that travelers' perceptions of public transit travel times are likely a function of whether they choose public transit or not. The frequency with which public transit is chosen may also affect perceptions (for further information see Chapter 2).

One may also hypothesize that travelers' perceptions of public transit walk and wait time are a function of the actual time values. Clark (1982) developed travel time perception models for different mode users, each based solely on the actual travel time values. The basis of each model was Steven's Law, a model used by psychologists to describe the relationship between the perceived magnitude of stimuli and their actual magnitude. The Steven's Law model assumes that perceptions of measured values vary exponentially. It has the following form,

$$PV=a(AV)^b$$

where: PV = perceived value

AV = actual value

a,b = constants

Clark (1982) determined coefficient values for each of the travel modes investigated. Some of his results are tabulated below.

	a	b	b<1?
AUTOMOBILE DRIVERS' PERCEPTIONS OF:			
1) CAR TRAVEL TIMES	2.61	0.604	YES
2) ALTERNATIVE MODE'S TRAVEL TIME	2.01	0.645	YES
BUS PASSENGERS' PERCEPTIONS OF:			
1) BUS TRAVEL TIMES	3.70	0.645	YES
2) ALTERNATIVE MODE'S TRAVEL TIME	2.88	0.536	YES

The "YES" in the b<1? column indicates that the "b" coefficient is less than 1 at the 1 percent level of significance. This result implies that the relationships between the perceived and measured times are non-linear.

As was argued in Chapter 4, persons with better public transit service may have better attitudes towards public transit and these attitudes may be reflected in their perceptions of public transit attributes. Since the composite utility measures the overall satisfaction an individual associates with his public transit alternatives, this variable may also explain traveler perception behaviour.

Socioeconomic characteristics might also explain travelers' perceptions of public transit attributes. O'Farrell and Markham (1974) suggest individual characteristics such as "age, social class and value orientations may be useful in attempting to explain distorted perceptions" (O'Farrell and Markham, 1974, p. 79). Equally relevant characteristics may be gender, income and employment type.

Trip purpose may also influence travelers' perceptions of public transit walk and wait times. The influence of this factor was not investigated in this research as the 1983 Morning Commuter Survey investigated characteristics of travelers' home-to-work trips only.

Therefore, it is proposed that travelers' perceptions of public transit walk times and wait time values are described by the following function,

$$P=f(x,s,u)$$

where: P = perceived attribute values

x = actual attribute values and composite utility values

s = socioeconomic characteristics

u = usual mode choice

This function was the basis for the linear regression models developed in this chapter.

7.2 Regression Analysis Theory

For this research, a regression model of the following form was assumed (Johnson and Wichern, 1982):

$$Y_i = \beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+....+\beta_k X_{ki}+....+\beta_p X_{pi}+\epsilon_i$$

where: Y_i = estimate of the dependent variable for case .

X_{ki} = value of the k^{th} predictor variable for case i

β_k = predictor variable weightings

ϵ_i = error term

This model states that, for each observation, the value of the dependent variable, Y, is equal to a linear function of the predictor variables $X_1, X_2, ...X_p$, plus a random error ϵ. In this analysis, the dependent variables were the travelers' perceived walk and wait times and the predictor variables were the actual travel time values, the travelers' socioeconomic characteristics and their usual mode choice.

The method of least squares estimates the regression coefficients, β_k, such that the resulting linear model best fits the observed data. Model goodness of fit is measured by the coefficient of determination, R^2. This coefficient measures the proportion of variation in the data that is explained by the model.

The use of linear regression requires the following assumptions (Norusis, 1982):

1. Values of the independent variable(s) are fixed.
2. Values of the dependent variable(s), for any value of the independent variable(s), are normally distributed with a constant variance.
3. The model error is normally distributed with a mean of zero and constant variance, and does not depend on the values assumed by the independent variable(s).

The analysis required for verification of assumptions 3 and 4 was not conducted. It was felt that the effort required for this analysis was beyond the scope of the intent of this linear regression analysis.

The regression analysis was performed using the SPSS-X package available on the University of Alberta mainframe computer. The "stepwise" option offered by the SPSS-X package was employed. This option allows the user to have the specified predictor variables added to the regression equation in a stepwise manner based on their ability to explain variation in the dependent variable; best predictors being added first, poorer predictors being added later. Variable inclusion criteria are defined in the program. The regression procedure terminates when the variables not yet included in the regression equation fail to meet these criteria. For further information on the stepwise option see the manual, "SPSS Introductory and Basic Statistics and

Operations" written by Norusis (1982).

7.3 Linear Regression Analysis

Two stepwise multivariate regression analyses were completed using the same eleven variables used for the development of the six cluster solution. The linear regression results are summarized below. Appendix G contains the SPSS-X program output.

MULTIVARIATE LINEAR REGRESSION RESULTS

REGRESSION 1:

REGRESSION EQUATION: PWALK=0.008AWALK-1.208MODE+0.505GENDER
+1.686

STEPWISE RESULTS:

STEP	VARIABLE ADDED	R^2
1	AWALK	.2534
2	MODE	.2845
3	GENDER	.2917

REGRESSION 2:

REGRESSION EQUATON: PWAIT=-2.887MODE+6.695

STEPWISE RESULTS:

STEP	VARIABLE ADDED	R^2
1	MODE	.1059

7.3.1 Model Goodness of Fit

Examination of the R^2 values for these models indicates that the perceived walk time model is somewhat better than the perceived wait time model. However, with R^2 values of 0.2917 and 0.1059, respectively, both are considered poor models. Possible explanations for such poor model fit are:

1. There is no relationship between the dependent variables and the hypothesized predictor variables.
2. The relationships between the dependent and predictor variables are not linear.

The results from the Fuzzy Cluster analysis suggest that the variables investigated in this research do, in fact, influence travelers' perceptions of their public transit walk and wait times. Explanation 2, therefore, is the likely explanation for the poor model fit. Further analysis of this aspect of the linear regression analysis is considered beyond the scope of the intent of this analysis.

7.3.2 Results

Recognizing that the regression results were poor, a brief examination of these results was still made.

The perceived walk time model suggests that a traveler's perception of his public transport walk time is a function of the actual walk distance, his usual mode choice and his gender. The actual walk distance appears to be the best predictor variable, of those included in this study, accounting for approximately 25.3% of the model variation.

The positive regression coefficient indicates, as one would expect, that an individual's perceived walk time tends to increase with an increase in actual walking distance. The next best predictor variable is the mode variable. The negative regression coefficient is also a logical one, indicating that the perceptions of public transit users (mode=1) are 1.2 minutes less, on average, than are the perceptions of private auto users (mode=0). The last predictor variable in the perceived walk time model is the gender variable. The positive regression coefficient model for this variable suggests that females (gender=2) perceive, on average, their public transit walk times 0.5 minutes longer than do men (gender=1).

The perceived wait time model suggests that travelers' perceptions of public transit wait time are a linear function of their regular mode choice only. The mode variable accounts for all 11% of the data variation explained by the model. It's negative coefficient indicates that public transit users estimate their public transit wait time 2.89 minutes less, on average, than do private auto users.

7.4 Model Comparisons

Because the same effort was not spent on the linear regression analysis as was spent on the Fuzzy Cluster approach, no direct comparison of the models' results were made. With regard to the effort required to run each of the models, it is concluded that the regression analysis was certainly the easier method to employ. This was largely due to the stepwise option available in the SPSS-X package. It is interesting to note that during the FCM analysis a "more-or-less" stepwise approach was attempted with the inclusion of variables being based on the variable correlation results with some subjective judgements. The availability of a stepwise option in the Fuzzy Cluster Method would have been very helpful in this research.

8. SUMMARY OF CONCLUSIONS AND RECOMMENDATIONS

The original objective of this research was to investigate the influence of socioeconomic and mode choice characteristics on travelers' perceptions of the walk times from their homes to their public transit stops and the wait times at these stops. Because of the perception component, a degree of "natural" imprecision was introduced into the research which led to the consideration of a "non-traditional" means of analysis - the Fuzzy Cluster Method. This method of analysis was chosen specifically because of its ability to represent, mathematically, imprecise phenomena. It was hypothesized that the result of this analysis would be clusters of perceptually homogenous travelers with regard to public transit travel time components. Recognizing the importance of travelers' perceptions of public transit walk and wait times in mode choice, it was further hypothesized that these clusters would essentially represent different public transit market segments. The primary objective of this research was therefore stated,

> "to use travelers' perceptions of public transport walk and wait time, their socioeconomic characteristics, and their usual mode choice to develop public transportation market segments"

Two secondary research objectives were also defined. They were:

1. To introduce Fuzzy Set Theory to transportation engineering practice in Edmonton, Canada.
2. To provide insight into travelers' perceptions of

time related public transit attributes.

The information required for this research was primarily derived from the "1983 Morning Commuter Survey" database. This disaggregate database was employed because it contained perceived and actual public transit data, socioeconomic data and mode choice information for morning peak hour commuters in Edmonton, Canada.

This chapter outlines the main results and conclusions that arose from this research, evaluates particular aspects of the research method, and provides a practical example of how the results of the cluster analysis may be used for development of public transit market strategies.

8.1 Main Conclusions

8.1.1 Evaluation of the Research Method

The perceptual component of this research provided an ideal circumstance for the investigation of the Fuzzy Cluster's potential for analysis of a practical transportation problem. To the author's knowledge, the Fuzzy Cluster Method had not previously been used for research related to travelers' time perceptions. As a result, some problems were encountered. The following discussion evaluates specific aspects of the research procedure, describes problems which were encountered and evaluates the resulting assumptions which were made.

1. The decision to use Fuzzy Cluster analysis for this

research was sound. Based on the nature of human perceptions, it was argued that one could not expect to categorize individuals into "crisply" defined groups but rather could expect some relatively stable yet "imprecise" patterns to exist. Because of this imprecision, the Fuzzy Cluster Method was considered an appropriate method of analysis.

Fuzzy clusters were, in fact, generated in this research. This conclusion is based on the following observations:

a. the effect of the weighting exponent value on the cluster results. Cluster validity measures indicated that there was no cluster structure in the data when a weighting exponent of 2.0 was used. However, with a reduced exponent value of 1.25 cluster structure was identified. It was hypothesized that this was a result of the presence of not-so-well separated (fuzzy) clusters; essentially, clusters that had to be truly fuzzy in order to be identified only by the lower exponent value which forced the objects (travelers) into crisper membership assignments.

b. the cluster splitting trends (Figure 6.2). It was observed that as the number of dimensions increased so too did the "best" number of clusters increase. Again, it was hypothesized that this was a result of not-so-well separated clusters in the database; clusters that were defined only as more information was made available for the clustering procedure.

It is also worth noting that because of the continuous nature of the Fuzzy Cluster objective function, it is differentiable with regard to the independent variables μ_{ik} and v_i. A minimum function value therefore exists for which optimal membership and cluster centre conditions are defined. Mathematically, this is advantageous in comparison to the crisp cluster procedure for which the objective function is discontinuous and, as a result, the continuous function definition of local minima does not apply.

2. The results of the cluster analysis appear to be intuitively correct. To the author's knowledge, no statistical measure of the reasonableness of a Fuzzy Cluster solution exists. In this research solution reasonableness was determined from: (1) examination of the cluster characteristics, both by the average cluster values and cluster cores and (2) examination of the cluster splitting trends.

   From the cluster splitting trends (Figure 6.2) it was observed that the "new" clusters in the higher dimension solutions were created through combining and splitting clusters of the previous solutions. It seems logical to conclude that the presence of previously defined clusters in the "new" solutions suggests stability for these particular clusters in the data.

3. The cluster size and cluster entropy measures developed for this research have provided a valid and useful method for measuring a cluster's size and fuzziness. Results obtained with these measure were verified by both the membership distribution graphs and the cluster core results (Chapter 6).
4. The determination of an appropriate weighting exponent was the largest obstacle encountered in the Fuzzy Cluster analysis procedure. Although the literature provided some guidance for a reasonable value, the final value that was used was attained, more-or-less, through trial-and-error. It is felt that in order for Fuzzy Cluster analysis to be considered a viable analytical approach for future practical applications, further analysis of appropriate weighting values must be undertaken.
5. The lack of a standard error measure made it impossible to test whether differences in cluster centre values were statistically significant. To the knowledge of the author, no standard deviation measure for the cluster centre values exists. The existence of such a statistic would certainly have been beneficial in this research.

6. The large number of iterations required for convergence of the Fuzzy C-Means program made this analysis demanding with regard to program running time (see Table 5.3). As was discussed in Section 5.2.3, the unusually large number of iterations required for the program's convergence was likely due to (1) the number of dimensions considered in this research, and (2) the complicated nature of the problem.

7. Regarding the assumptions made for the actual walk and wait time calculations it is concluded that:

    a. the assumption of a 1.2 m/s average walking speed was a reasonable one.

    b. the assumption that an individual's wait time was equal to one-half of the public transit service frequency was poor for the 1983 Edmonton situation (headways were generally greater than 10 minutes). It is highly unlikely that most individuals underestimated their public transit wait time as was suggested by the actual wait time results. More likely, the underlying assumption that people were arriving randomly at all public transit stops was incorrect. Had the author assumed a more systematic arrival pattern, a shorter, and seemingly more accurate predicted wait time would have resulted. It is stressed, however, that because the same actual wait time model was used for all individuals, and provided a means of comparison for the perceived travel time values only, the results obtained with these values are considered valid.

8.1.2 Travelers' Perceptions

An associated benefit of this research was insight into how travelers' perceive two specific public transit attributes: public transit walk times and wait times. With regard to this aspect of the research, it is concluded that travelers' perceptions of public transit walk times and public transit wait times are influenced by the actual magnitudes of these variables, socioeconomic characteristics, mode choice and the frequency with which this mode is usually used. The socioeconomic characteristics specifically investigated in this research were age, gender, income and job type.

This conclusion is based on the manner in which clusters were created as variables were added in the cluster analysis procedure (see Fig. 6.2). The least number of variables considered in any one Fuzzy C-Means program run was six. Past research had indicated that car drivers' and public transit users' perceptions of public transit attributes were in fact different. Therefore, perceived and actual public transit time variables as well as the mode variable were chosen as input dimensions for this program run. The composite utility value was also included because of its relatively high correlation with many of the research variables. With these variables two clusters were identified: one cluster of public transit users and one cluster of auto drivers.

Gradually the other research variables were added into the cluster analysis procedure. As variables were added, more clusters were identified until all eleven variables were considered in the analysis. For the eleven dimension analysis six clusters were found to best represent the cluster structure in the data. Further analysis of the characteristics of the clusters concentrated on this six cluster solution because:

1. this solution seemed to have evolved from the other cluster solutions with less clusters, and
2. all eleven variables were represented in this solution.

The socioeconomic and mode choice characteristics of the clusters identified in the six cluster solution are summarized below:

Cluster 1: is characterized by relatively low income (avg. income 18,300 $/yr), fairly young (avg. age 26.7 yrs), female public transit users who, on average, use public transit less often than do the individuals who belong to the other public transit using clusters.

Cluster 2: contains low income (avg. 15,900 $/yr), older (avg. 45.8 yrs) female public transit users.

Cluster 3: contains middle-aged (avg. 30.8 yrs) male and female auto users, who earn, on average 21,800 $/yr.

Cluster 4: is characterized by middle-aged (avg. 34.3 yrs) male public transit users, who earn, on average 26,650 $/yr.

Cluster 5: contains slightly older (avg. 38.5 years), males and females (mostly males), who have, on average, the highest incomes of the individuals in the sample data (avg. cluster income 36,200 $/yr). Mode usage for this cluster is almost split evenly between public transit and private vehicle.

Cluster 6: is characterized by low income (avg. 15,950 $/yr), young (avg. 24.0 yrs), female public transit users, who use public transit more often than do those females of cluster 1.

Perceptual accuracy was assessed relative to actual (estimated) transit time values. The six cluster solution had the following perceptual characteristics:

1. Clusters 4,5, and 6 had the lowest underestimations of public transit walk and wait times. This was deemed a logical result as both clusters 4 and 6 contained individuals who used public transit

regularly for their morning home-to-work trip. Somewhat surprising were the accurate time estimations for the high income individuals of cluster 5, as only one-half of these individuals used public transit regularly.

2. Clusters 1,2, and 3 had the highest overestimations of public transit walk and wait times. For the less frequent public transit users of cluster 1 and the auto drivers of cluster 3, this was a logical result. This same result, however, was surprising for the regular transit users of cluster 2.

T-test results indicated that differences between clusters with extremely different perceptual characteristics were significant at the 90% level of significance. Differences between perceptually similar clusters were not significant at this level. For further detail see Figures 6.17 and 6.18.

8.1.3 Linear Regression Results

For comparison, two multivariate linear regression models were developed; one model to represent travelers' perceptions of public transit walk times and one to represent travelers' perceptions of public transit wait times. The perceived walk time model indicated that travelers' perceptions of public transit walk time was influenced by actual walk distances, usual mode choice and gender. The only variable identified as significant in the wait time model was usual mode choice. Both models were, however, considered poor models with R^2 values of 0.2917 and 0.1059, respectively. Because the Fuzzy Cluster results identified perceptual trends amongst the travelers, the poor linear regression results were attributed to the degree of complexity of human perceptions and the simplicity of the linear regression model. The regression analysis was not rigorously pursued in this project.

8.2 Practical Implications of the Research Results

The following section provides a specific example in order to illustrate how the market segmentation results derived in this research may be used for developing market strategies targetted at specific groups of travelers.

This example examines the transit market potential for cluster 3. This cluster contains middle-income (avg. 21,800 $/yr), middle-aged (avg. 30.8 yrs), male and female car users. What is particularly noteworthy about this cluster is that, relative to the other clusters, individuals from this cluster have the highest overestimations of their public transit walk and wait times. These perceptions are reflected in the average perception difference and perception ratio values for this cluster (Appendix B) and the cluster core walk and wait time perception distributions (Appendix D). The high walk and wait time estimations suggest a lack of public transit service knowledge.

From a marketing standpoint, then, it is important to improve these travelers' perceptions of public transit service before consideration of actual service improvements are made. The following discussion addresses specific considerations for meaningful market segmentation development.

The socioeconomic and mode choice characteristics of the travelers in this market segment are provided by the Fuzzy Cluster results. Information regarding the substantiality of this market can also be derived from these research results. The cluster size measure indicates that cluster 3 is a reasonably large cluster (approximately 15% of the sample data). The total number of travelers in this market segment could actually be approximated using this percentage, the total sample size, and an estimate of the number of morning downtown commuters in Edmonton. Equally important with regard to the substantiality of this market is the degree of fuzziness of this cluster. Cluster 3 was the fuzziest cluster identified in this analysis. Arguing that cluster members with lower degrees of membership, i.e. those with a lower degree of cluster "loyalty", may be more susceptible to marketing strategies aimed at changing particular characteristics, reinforces the conclusion that cluster 3 is a good candidate for target marketing.

Many opportunities exist to focus marketing efforts on this market segment. For instance, since this cluster contains infrequent public transit users one might consider relaying public transit information through posted schedules and/or automatic telephone schedule information. Advertising should concentrate on service frequency information, typical travel times, etc.

8.3 Concluding Remarks

This pilot project has proven that Fuzzy Cluster Analysis has considerable merit as a tool for developing public transit market segments based on travelers' perceptions of two components of a public transit trip. In this research up to six clusters, each with unique socioeconomic, mode choice, and perceptual characteristics were identified.

The procedure was not, however, without shortcomings. They included:

1. no benchmark cluster validity measure to indicate the reasonableness of a cluster solution.
2. lack of information regarding appropriate weighting values.
3. no cluster centre standard error measure.

It is felt that in order for Fuzzy Cluster analysis to be considered a viable analytical approach for future practical transportation applications, these limitations must be addressed.

References

ALGERS, S., S. HANSEN, and G. TEGNER. 1975. Role of Waiting Time, Comfort and Convenience in Modal Choice for Work Trip. Transportation Research Record 534:38-51.

BEN-AKIVA, M. and S. LERMAN. 1979. Disaggregate Travel and Mobility Choice Models and Measures of Accessibility. Behavioural Travel Modelling, eds. D. Hensher and P. Stopher. London: Croom Helm.

BEZDEK, J.C. 1974. Cluster Validity with Fuzzy Sets. Journal of Cybernetics 3(3):58-73.

BEZDEK, J..C. 1980. A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 2(1) (IEEE PAMI-2):1-8.

BEZDEK, J.C. 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press.

BEZDEK, J.C., R. EHRLICH, and W. FULL. 1984. FCM: The Fuzzy C-Means Clustering Algorithm. Computers and Geosciences 10(2-3):191-203.

BOVY, P.H.L. and G.R.M. Jansen. 1979. Travel Times for Disaggregate Travel Demand Modelling: A Discussion and a New Travel Time Model. New Developments in Modelling Travel Demand and Urban Systems, eds. G.R.M. Jansen, P.H.L. Bovy, J.P.J.M. Van Est, and F. Le Clerq. England: Saxon House.

BOX, G.E.P., W.G. HUNTER, and J.S. HUNTER. 1978. Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. New York: John Wiley & Sons.

CANADIAN URBAN TRANSIT ASSOCIATION (CUTA). 1985. Canadian Transit Handbook, Second Edition. Canadian Urban Transit Association and Roads and Transportation Association of Canada:22(1)-22(15).

CHAPMAN, R.A., H.E. GAULT, and I.A. JENKINS. 1976. Factors Affecting the Operation of Urban Bus Routes. University of Newcastle Upon Tyne Research Report No. 23.

CLARK, J.E. 1982. Modeling Travelers' Perceptions of Travel Time. Transportation Research Record 890:7-11.

COOPER, D.L. 1989. Transit Route Analysis Model (TRAM). M.Sc. Thesis, Department of Civil Engineering, University of Alberta, Canada.

DE LUCA, A. and S. TERMINI. 1972. A Definition of a Nonprobabilistic Entropy in the Setting of Fuzzy Sets Theory. Information and Control 20:301-312.

DOMENCICH, T.A. and D. MCFADDEN. 1975. Urban Travel Demand: A Behavioural Analysis. Amsterdam: North Holland.

DUNN, J.C. 1974. A Fuzzy Relative of the ISODATA Process and its use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics 3(3):32-57.

DUNN, J.D. 1974. Some Recent Investigations of a New Fuzzy Partitioning Algorithm and its Application to Pattern Classification Problems. Journal of Cybernetics 4(2):1-15.

ESOGBUE, A.O. 1986. Optimal Clustering of Fuzzy Data Via Fuzzy Dynamic Programming. Fuzzy Sets and Systems 18:283-298.

GEARY, R.C. 1936. The Distribution of Student's Ratio for Non-Normal Samples. Journal of The Royal Statistical Society 3:178-184.

GILBERT, G. and J. FOERSTER. 1977. The Importance of Attitudes in the Decision to Use Mass Transit. Transportation 6:321-332.

HEGGIE, I.G. 1976. Modal Choice and the Value of Travel Time. Oxford: Clarendon Press.

HENSHER, D.A. 1975. Perception and Commuter Modal Choice - An Hypothesis. Urban Studies 12:101-104.

HENSHER, D.A. 1976. Market Segmentation as a Mechanism in Allowing for Variability of Traveller Behaviour. Transportation 5(3):257-284.

HUNT, J.D. 1984. Description of the Morning Commuter Survey. Ph.D. Research Working Paper 7. Unpublished, Department of Civil Engineering, University of Alberta, Canada.

HUNT, J.D. 1988. Modelling Commuter Parking Location Choice and its Influence on Mode Choice. Ph.D. Dissertation, Churchill College, University of Cambridge, United Kingdom.

JOHNSON R.A. and D.W. WICHERN. 1982. Applied Multivariate Statistical Analysis. Englewood Cliffs, New Jersey: Prentice-Hall Inc.

MEYBURG, A.H. and W. BROG. 1981. Validity Problems in Empirical Analyses of Non-Home Activity Patterns. Transportation Research Record 807:46-50.

NICOLAIDIS, G.C. and R. DOBSON. 1975. Disaggregated Perceptions and Preferences in Transportation Planning. Transportation Research 9:279-295.

NICHOLSON, B.K. 1987. Pedestrian Walking Speed: A Comparison of Winter and Summer Conditions. M. Eng. Project, Department of Civil Engineering, University of Alberta, Canada.

NORUSIS, M.J. 1982. SPSS Introductory and Basic Statistics and Operations. New York: McGraw-Hill.

O'FARRELL, P.N. and J. MARKHAM. 1974. Commuter Perceptions of Public Transport Work Journeys. Environment and Planning A 6:79-100.

QUARMBY, D.A. 1967. Choice of Travel Mode for the Journey to Work: Some Findings. Journal of Transport Economics and Policy 1(3):273-301.

RECKER, W.W. and T.F. GOLOB. 1976. An Attitudinal Modal Choice Model. Transportation Research 10:299-310.

STATISTICAL RESEARCH LABORATORY, UNIVERSITY OF MICHIGAN. 1976. Elementary Statistics Using MIDAS. The University of Michigan: The Statistical Laboratory.

STOPHER, P.R. 1977. Development of Market Segments of Destination Choice. Transportation Research Record 649:14-22.

STOPHER, P.R. and A.H. Meyburg. 1975. Behavioral Travel Demand Models. Behavioral Travel-Demand Models, eds. P.R. Stopher and A.H. Meyburg. Lexington, Massachusetts: Lexington Books, D.C. Heath and Company.

STOPHER, P.R. and A.H. MEYBURG. 1979. Survey Sampling and Multivariate Analysis for Social Scientists and Engineers. Lexington, Massachusetts: Lexington Books, D.C. Heath and Company.

WATSON, P.L. 1971. Problems Associated with Time and Cost Data Used in Travel Choice Modeling and Valuation of Time. Highway Research Record 369:148-158.

WINDHAM, M.P. 1981. Cluster Validity for Fuzzy Clustering Algorithms. Fuzzy Sets and Systems 5:177-185.

WINDHAM, M.P. 1983. Geometrical Fuzzy Clustering Algorithms. Fuzzy Sets and Systems 10:271-279.

ZIMMERMAN, H.J. 1985. Fuzzy Set Theory - and Its
Applications. Boston: Kluwer-Nijhoff Publishing.

Appendix A
Fuzzy C-Means Program Source Code

```
C ______________________________________________________________________


C     COMPILER FORTGICG *
C     FUZZY C-MEANS PROGRAM: WILLIAM E. FULL, WICHITA STATE UNIVERSITY,
C                            WICHITA, KANSAS   67208  (DEPT. OF GEOLOGY)
C     THIS IS THE FCM (FUZZY C-MEANS) ROUTINE. THIS LISTING IS FOR A
C     IBM TYPE COMPUTER WITH A FORTRAN IV COMPILER. IT ADAPTS FOR ANY
C     FORTRAN COMPILER WITH MODIFICATIONS SET AT THE USER SITE.
C
C     REFERENCE: "PATTERN RECOGNITION WITH FUZZY OBJECTIVE FUNCTIONS,"
C                JAMES BEZDEK, PLENUM, NEW YORK, 1981.
C
C          NOTE: THIS PROGRAM INITIALIZES THE MEMBERSHIP MATRIX USING
C                THE RANDOM=.7731 DEFINITION.
C
C
C     DESCRIPTION OF OPERATING VARIABLES:
C     I. INPUT VARIABLES (FROM FILE 5)
C         CARD 1:
C          TITLE(20)........80 CHARACTER HEADING
C         CARD 2:
C          FMT(20)..........FORTRAN FORMAT (CONTAINED IN PARENTHESIS)
C                           DESCRIBING THE INPUT FORMAT FOR THE RAW DATA
C                           UP TO 80 CHARACTERS MAY BE USED
C         CARD 3:
C          COL 1: ICON......DISTANCE MEASURE TO BE USED. IF:
C                           ICON=1 USE EUCLIDEAN NORM
C                           ICON=2 USE DIAGONAL NORM
C                           ICON=3 USE MAHALANOBIS NORM
C          COLS 2-7: QQ.....WEIGHTING EXPONENT FOR FCM
C          COLS 8-9: ND.....NUMBER OF FEATURES PER INPUT VECTOR
C          COLS 10-11:KBEGIN.STARTING NUMBER OF CLUSTERS
C          COLS 12-13:KCEASE.FINISHING NUMBER OF CLUSTERS (NOTE: KBEGIN
C                           MUST BE LESS THAN OR EQUAL TO KCEASE)
C         CARD 4 ON:
C          Y(NS,ND).........FEATURE VECTORS, INPUT ROW-WISE
C     II. INTERNAL VARIABLES
C          NS...............NUMBER OF DATA VECTORS
C          EPS..............MAXIMUM MEMBERRSHIP ERROR AT CONVERGENCE
C          NC...............CURRENT NUMBER OF CLUSTERS
C          LMAX.............MAXIMUM NUMBER OF ITERATIONS WITHOUT
C                           CONVERGENCE
C          FM(ND)...........SAMPLE MEAN VECTOR
C          FVAR(ND).........VECTOR OF MARGINAL VARIANCES
C          CC(ND,ND)........SCALING MATRIX
C          AA(ND,ND)........SAMPLE COVARIANCE MATRIX
C          AI(ND,ND)........INVERSE OF SAMPLE COVARIANCE MATRIX
C          BB(ND)...........DUMMY HOLDING MATRIX
C          CCC(ND)..........DUMMY HOLDING MATRIX
C          ST(ND,ND)........DUMMY HOLDING MATRIX FOR AA
C          CM(ND,ND)........CM=AA*(AA INVERSE)
C          U(NC,NS).........MEMBERSHIP MATRIX
C          W(NC,NS).........UPDATED MEMBERSHIP MATRIX
C          V(NC,ND).........CLUSTER CENTERS
C          ITT(NC)..........DUMMY HOLDING MATRIX
C          H(NC)............ENTROPY MATRIX
C          VJM(NC)..........PAYOFF MATRIX
```

```
C          F(NC)............MATRIX OF PARTITION COEFFICIENTS
C          DIF(NC)..........MATRIX OF ENTROPY BOUNDS
C
C
C*****************************************************************
      DIMENSION FM(15),FVAR(15),F(20)
      DIMENSION BB(15),CCC(15),H(20),DIF(20),ITT(20)
      DIMENSION Y(750,15),U(20,750),W(20,750)
      DIMENSION AA(15,15),AI(15,15)
      DIMENSION CC(15,15),CM(15,15),ST(15,15)
      DIMENSION V(20,15),VJM(20)
      DIMENSION FMT(20),TITLE(20)
      READ(5,1458) (TITLE(I),I=1,20)
1458  FORMAT(20A4)
      READ(5,12321) (FMT(I),I=1,20)
12321 FORMAT(20A4)
C-----------------------------------------------------------------
C     CONTROL PARAMETERS.
C-----------------------------------------------------------------
      EPS=.01
      NS=1
      LMAX=50
C-----------------------------------------------------------------
C     READ FEATURE VECTORS (Y(I,J)).
C-----------------------------------------------------------------
      READ(5,2021) ICON,QQ,ND,KBEGIN,KCEASE
2021  FORMAT(I1,F6.3,3I2)
      WRITE(6,410)
410   FORMAT(///1H ,'*** *** BEGIN FUZZY C-MEANS OUTPUT *** ***')
      WRITE(6,1459) (TITLE(III),III=1,20)
1459  FORMAT(10X,20A4///)
1     READ(5,FMT,END=3)(Y(NS,J),J=1,ND)
      WRITE(6,12738)(Y(NS,J),J=1,ND)
12738 FORMAT(2(10X,10(F7.2,1X)/))
      NS=NS+1
      GO TO 1
3     NS=NS-1
      NDIM=ND
      NSAMP=NS
      WRITE(6,11111) NSAMP
11111 FORMAT(10X,'NUMBER OF SAMPLES = ',I5)
      ANSAMP=NSAMP
C-----------------------------------------------------------------
C     SCALED NORM REQUIRED IN STATEMENTS 31 AND 33.
C     CALCULATION OF SCALING MATRIX FOLLOWS.
C     FEATURE MEANS.
C-----------------------------------------------------------------
      DO 350 I=1,NDIM
      FM(I)=0.
      DO 351 J=1,NSAMP
351   FM(I)=FM(I)+Y(J,I)
350   FM(I)=FM(I)/ANSAMP
C-----------------------------------------------------------------
C     FEATURE VARIANCES.
C-----------------------------------------------------------------
      DO 352 I=1,NDIM
      FVAR(I)=0.
      DO 353 J=1,NSAMP
```

```
353   FVAR(I)=FVAR(I)+((Y(J,I)-FM(I))**2)
352   FVAR(I)=FVAR(I)/ANSAMP
      IF(ICON-1)380,380,382
380   DO 381 I=1,NDIM
      DO 381 J=1,NDIM
381   CC(I,J)=0.
      DO 370 I=1,NDIM
370   CC(I,I)=1.
      GO TO 390
382   IF(ICON-2)384,384,386
384   DO 385 I=1,NDIM
      DO 385 J=1,NDIM
385   CC(I,J)=0.
      DO 371 I=1,NDIM
371   CC(I,I)=1./FVAR(I)
      GO TO 390
386   DO 360 I=1,NDIM
      DO 360 J=1,NDIM
      AA(I,J)=0.
      DO 361 K=1,NSAMP
361   AA(I,J)=AA(I,J)+((Y(K,I)-FM(I))*(Y(K,J)-FM(J)))
360   AA(I,J)=AA(I,J)/ANSAMP
      DO 550 I=1,NDIM
      DO 550 J=1,NDIM
550   ST(I,J)=AA(I,J)
C-----------------------------------------------------------------------
C     INVERSION OF COVARIANCE MATRIX AA TO AI
C-----------------------------------------------------------------------
      NN=NDIM-1
      AA(1,1)=1./AA(1,1)
      DO 500 M=1,NN
      K=M+1
      DO 501 I=1,M
      BB(I)=0.
      DO 501 J=1,M
501   BB(I)=BB(I)+AA(I,J)*AA(J,K)
      D=0.
      DO 502 I=1,M
502   D=D+AA(K,I)*BB(I)
      D=-D+AA(K,K)
      AA(K,K)=1./D
      DO 503 I=1,M
503   AA(I,K)=-BB(I)*AA(K,K)
      DO 504 J=1,M
      CCC(J)=0.
      DO 504 I=1,M
504   CCC(J)=CCC(J)+AA(K,I)*AA(I,J)
      DO 505 J=1,M
505   AA(K,J)=-CCC(J)*AA(K,K)
      DO 500 I=1,M
      DO 500 J=1,M
500   AA(I,J)=AA(I,J)-BB(I)*AA(K,J)
      DO 520 I=1,NDIM
      DO 520 J=1,NDIM
520   AI(I,J)=AA(I,J)
      DO 387 I=1,NDIM
      DO 387 J=1,NDIM
387   CC(I,J)=AI(I,J)
```

```
C----------------------------------------------------------------------
C       CHECK INVERSE AA*AI=I
C----------------------------------------------------------------------
        DO 530 I=1,NDIM
        DO 530 J=1,NDIM
        CM(I,J)=0.
        DO 530 K=1,NDIM
530     CM(I,J)=CM(I,J)+ST(I,K)*AI(K,J)
        WRITE(6,531)
531     FORMAT(' ',//,' CHECK MATRIX A1*AA=1, THE IDENTITY'//)
        DO 532 I=1,NDIM
532     WRITE (6,533) (CM(I,J),J=1,NDIM)
533     FORMAT(10X,20F6.2)
390     WRITE(6,1460) (TITLE(III),III=1,20)
1460    FORMAT('1',10X,20A4///)
        WRITE(6,420)
420     FORMAT(' ',///,15X,'SCALING MATRIX CC',///)
        DO 421 I=1,NDIM
421     WRITE(6,422) (CC(I,J),J=1,NDIM)
422     FORMAT(5X,10(F10.1,1X)/5X,10(F10 1,1X)/)
        WRITE(6,425)
425     FORMAT(/////)
C----------------------------------------------------------------------
C       QQ IS THE BASIC EXPONENT FOR FUZZY ISODATA.
C----------------------------------------------------------------------
        PP=(1./(QQ-1.))
        DO 55555 NCLUS=KBEGIN,KCEASE
        WRITE(6,1460) (TITLE(III),III=1,20)
        WRITE(6,499) NCLUS,ICON,QQ
499     FORMAT(' ',' NUMBER OF CLUSTERS = ',I3,5X,' ICON = ',I3,5X,
       C'EXPONENT = ',F4.2,//)
        IT=1
C----------------------------------------------------------------------
C       RANDOM INITIAL GUESS FOR U(I,J)
C       THE RANDOM GENERATOR SUBROUTINE RANDU FROM THE IBM SCIENTIFIC
C       SUBROUTINE PACKAGE (SSP) IS USED AND IS CALLED FROM AN EXTERNAL
C       LIBRARY. OTHER GENERATORS THAT PRODUCE VALUES ON THE INTERVAL
C       ZERO TO ONE CAN BE USED.
C----------------------------------------------------------------------
        RANDOM=.7731
        IX=1
        NCLUS1=NCLUS-1
        DO 1100 K=1,NSAMP
        S=1.0
        DO 1101 I=1,NCLUS1
C       CALL RANDU(IX,IY,RANDOM)
        RANDOM=RANDOM/2.
        IX=IY
        ANC=NCLUS-I
        U(I,K)=S*(1.0-RANDOM**(1.0/ANC))
1101    S=S-U(I,K)
1100    U(NCLUS,K)=S
C----------------------------------------------------------------------
C       CALCULATION OF CLUSTER CENTERS V(I).
C----------------------------------------------------------------------
7000    DO 20 I=1,NCLUS
        DO 20 J=1,NDIM
        V(I,J)=0.
```

```
      D=O.
      DO 21 L=1,NSAMP
      V(I,J)=V(I,J)+((U(I,L)**QQ)*Y(L,J))
21    D=D+(U(I,L)**QQ)
20    V(I,J)=V(I,J)/D
C-----------------------------------------------------------------------
C     UPDATE MEMBERSHIP FUNCTIONS.
C-----------------------------------------------------------------------
6111  DO 38 I=1,NCLUS
      DO 38 J=1,NSAMP
      W(I,J)=O.
      A=O.
      DO 31 L=1,NDIM
      DO 31 M=1,NDIM
31    A=A+((Y(J,L)-V(I,L))*CC(L,M)*(Y(J,M)-V(I,M)))
      A=1./(A**PP)
      SUM=O.
      DO 32 N=1,NCLUS
      C=O.
      DO 33 L=1,NDIM
      DO 33 M=1,NDIM
33    C=C+((Y(J,L)-V(N,L))*CC(L,M)*(Y(J,M)-V(N,M)))
      C=1./(C**PP)
32    SUM=SUM+C
      W(I,J)=A/SUM
38    CONTINUE
C-----------------------------------------------------------------------
C     ERROR CRITERIA AND CUTOFFS.
C-----------------------------------------------------------------------
9000  ERRMAX=O.
      DO 40 I=1,NCLUS
      DO 40 J=1,NSAMP
      ERR=ABS(U(I,J)-W(I,J))
      IF(ERR.GT.ERRMAX) ERRMAX=ERR
40    CONTINUE
      WRITE(6,400) IT,ERRMAX,NCLUS
400   FORMAT(1H ,'ITERATION = ',I4,5X,'MAXIMUM ERROR = ',F10.4,
     110X,'NUMBER OF CLUSTERS = ',I4)
      DO 42 I=1,NCLUS
      DO 42 J=1,NSAMP
42    U(I,J)=W(I,J)
      IF(ERRMAX.LE.EPS) GO TO 6000
43    IT=IT+1
      IF(IT-LMAX) 7000,7000,6000
C-----------------------------------------------------------------------
C     CALCULATION OF CLUSTER VALIDITY STATISTICS F, H, 1-E
C-----------------------------------------------------------------------
6000  ITT(NCLUS)=IT
      F(NCLUS)=0.0
      H(NCLUS)=0.0
      DO 100 I=1,NCLUS
      DO 100 K=1,NSAMP
      AU=U(I,K)
      F(NCLUS)=F(NCLUS)+AU**2/ANSAMP
      IF (AU) 100,100,101
101   H(NCLUS)=H(NCLUS)-AU*ALOG(AU)/ANSAMP
100   CONTINUE
      DIF(NCLUS)=1.0-F(NCLUS)
```

```
C-----------------------------------------------------------------------
C       CALCULATION OF OBJECTIVE FUNCTI N
C-----------------------------------------------------------------------
        A=0.
        DO 80 I=1,NCLUS
        DO 80 J=1,NSAMP
        DIST=0.
        DO 81 L=1,NDIM
        DO 81 M=1,NDIM
81      DIST=DIST+((Y(J,L)-V(I,L))*CC(L,M)*(Y(J,M)-V(I,M)))
        A=A+((U(I,J)**QQ)*DIST)
80      VJM(NCLUS)=A
C-----------------------------------------------------------------------
C       OUTPUT BLOCK FOR CURRENT NCLUS
C-----------------------------------------------------------------------
        WRITE(6,401)
401     FORMAT(' '///' FSTOP',7X,'1-FSTOP',5X,'ENTROPY',5X,'PAYOFF',5X,/)
        WRITE(6,699) F(NCLUS),DIF(NCLUS),H(NCLUS),VJM(NCLUS)
699     FORMAT(1H ,2(F6.3,4X),4X,F6.3,5X,E8.3)
        WRITE(6,59)
59      FORMAT(1X,100('-')//)
        WRITE(6,402)
402     FORMAT(///,15X,'CLUSTER CENTERS V(I,J)',///)
        DO 415 I=1,NCLUS
415     WRITE(6,404) (I,J,V(I,J),J=1,NDIM)
404     FORMAT(' I=',I3,3X,'J=',I3,3X,'V(I,J)=',F9.4)
405     FORMAT(1H ,7(F6.4,3X))
        WRITE(6,59)
        WRITE(6,406)
406     FORMAT(1H ,///,25X,'MEMBERSHIP FUNCTIONS',///)
        DO 407 J=1,NSAMP
407     WRITE(6,408) J,(U(I,J),I=1,NCLUS)
408     FORMAT(1H ,'J=',I3,5X,8(F6.4,3X))
54444 CONTINUE
55555 CONTINUE
C-----------------------------------------------------------------------
C       OUTPUT SUMMARY FOR ALL VALUES OF C
C-----------------------------------------------------------------------
        WRITE(6,450)
450     FORMAT('1',25X,'RUN SUMMARY')
        WRITE(6,460) NSAMP
460     FORMAT(' '///' NUMBER OF SUBJECTS N = ',I4)
        WRITE(6,461) NDIM
461     FORMAT(1HO,'NUMBER OF FEATURES NDIM = ',I4)
        WRITE(6,462) EPS
462     FORMAT(1HO,'MEMBERSHIP DEFECT BOUND EPS = ',F6.4)
        WRITE(6,464) ICON
464     FORMAT(1HO,'NORM THIS RUN ICON = ',I1)
        WRITE(6,465) QQ
465     FORMAT(1HO,'WEIGHTING EXPONENT M = ',F4.2)
        IF(IT.LE.49) GO TO 476
        WRITE(6,70107)
70107 FORMAT(' ','CONVERGENCE FLAG: UNABLE TO ACHIEVE SATISFACTORY CLUST
       1ERS AFTER 50 ITERATIONS.')
476     WRITE(6,466)
466     FORMAT(' '//' NO. OF CLUSTERS',3X,'PART. COEFF.',5X,
       C'LOWER BOUND',5X,'ENTROPY',5X,'NUMBER OF ITERATIONS')
        WRITE(6,467)
```

```
467   FORMAT(1HO,6X,'C',17X,'F',15X,'1-F',12X,'H',10X,'IT')
      DO 468 J=KBEGIN,KCEASE
468   WRITE(6,469) J,F(J),DIF(J),H(J),ITT(J)
469   FORMAT(1H ,6X,I2,14X,F6.3,11X,F6.3,7X,F6.3,8X,I4)
55556 CONTINUE
616   WRITE(6,411)
411   FORMAT(////1H ,'*** *** NORMAL END OF JOB *** ***')
      STOP
      END
      SUBROUTINE RANDU(IX,IY,YFL)
      IY=IX*65539
      IF(IY)5,6,6
    5 IY=IY+2147483647+1
    6 YFL=IY
      YFL=YFL*.4656613E-9
      RETURN
      END
```

Appendix B
Cluster Solutions and Membership Distribution Graphs

TWO CLUSTER SUMMARY

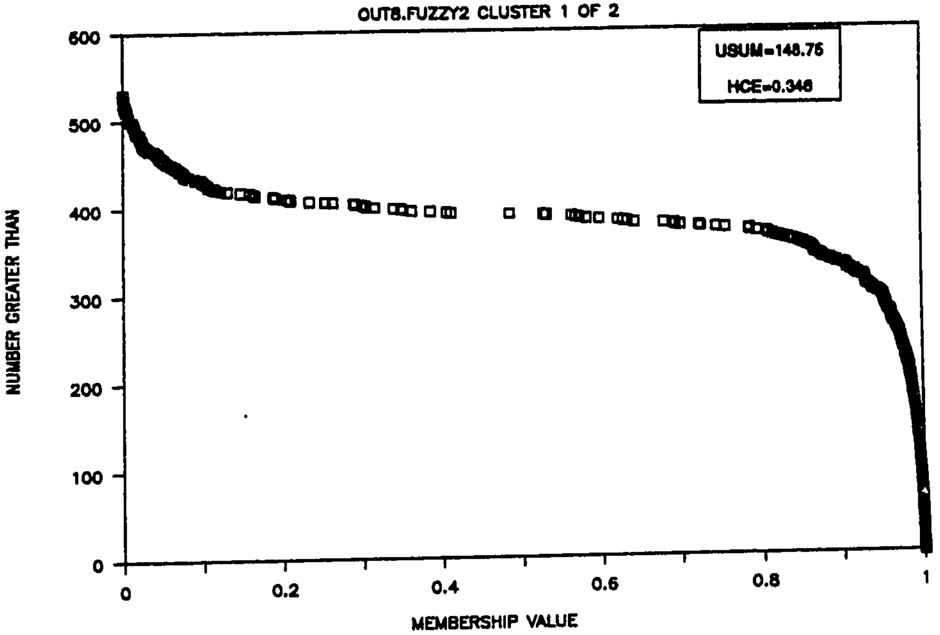
DIMENSION	CLUSTER	
	1	2
PWALK	3.344	4.259
PWAIT	3.885	6.197
AWALK	253.063	287.488
AWAIT	12.396	14.377
COMPU	-7.052	-8.561
MODE	0.979	0.090

PERCEPTION RATIOS (PERCEIVED/ACTUAL):

WALK	0.95	1.08
WAIT	0.63	0.86

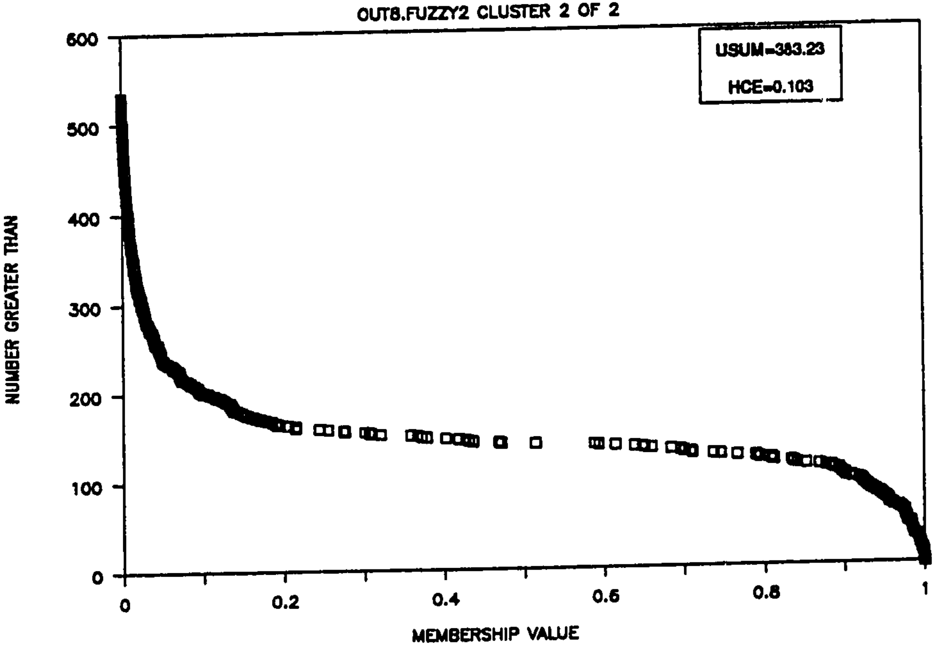
PERCEPTION DIFFERENCES (ACTUAL - PERCEIVED):

WALK (min)	0.17	-0.27
WAIT (min)	2.31	0.99



MEMBERSHIP DISTRIBUTION

OUT8.FUZZY2 CLUSTER 1 OF 2



MEMBERSHIP DISTRIBUTION

OUT8.FUZZY2 CLUSTER 2 OF 2

THREE CLUSTER SUMMARY

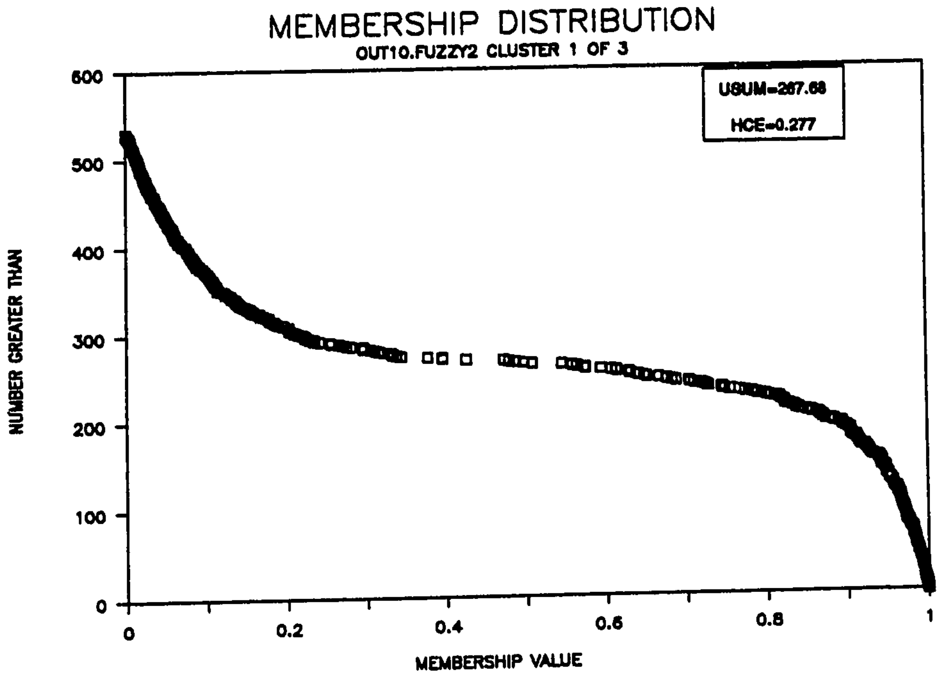
DIMENSION	CLUSTER		
	1	2	3
PWALK	3.381	4.049	3.483
PWAIT	4.073	5.906	3.927
AWALK	248.338	279.723	269.922
AWAIT	11.755	13.988	14.203
COMPU	-6.811	-8.469	-7.776
INCOME	6.251	8.068	8.745
GENDER	1.965	1.563	1.111
MODE	0.963	0.082	0.914

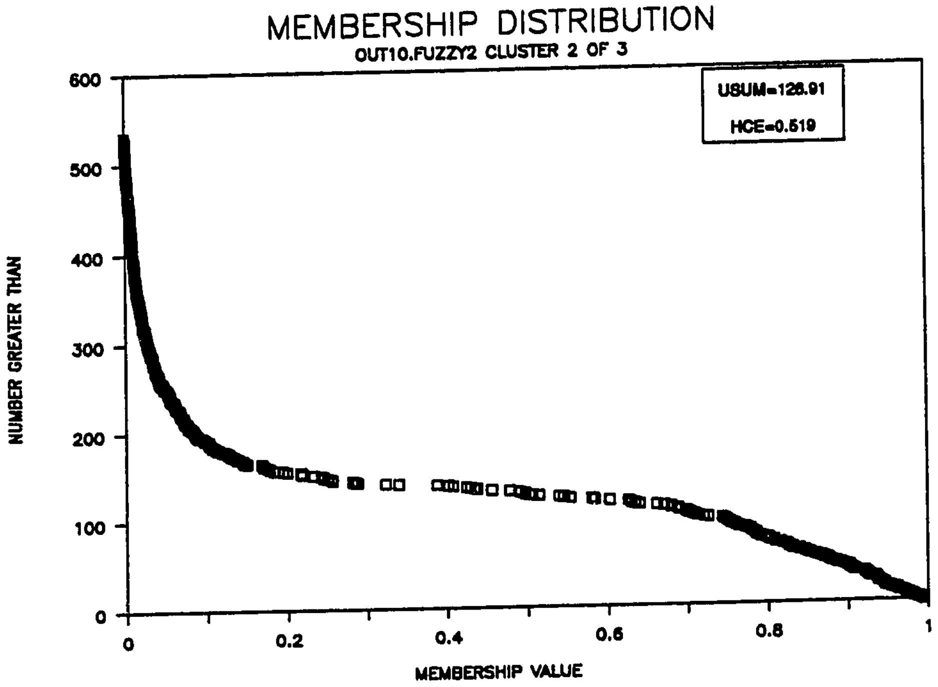
PERCEPTION RATIOS (PERCEIVED/ACTUAL):

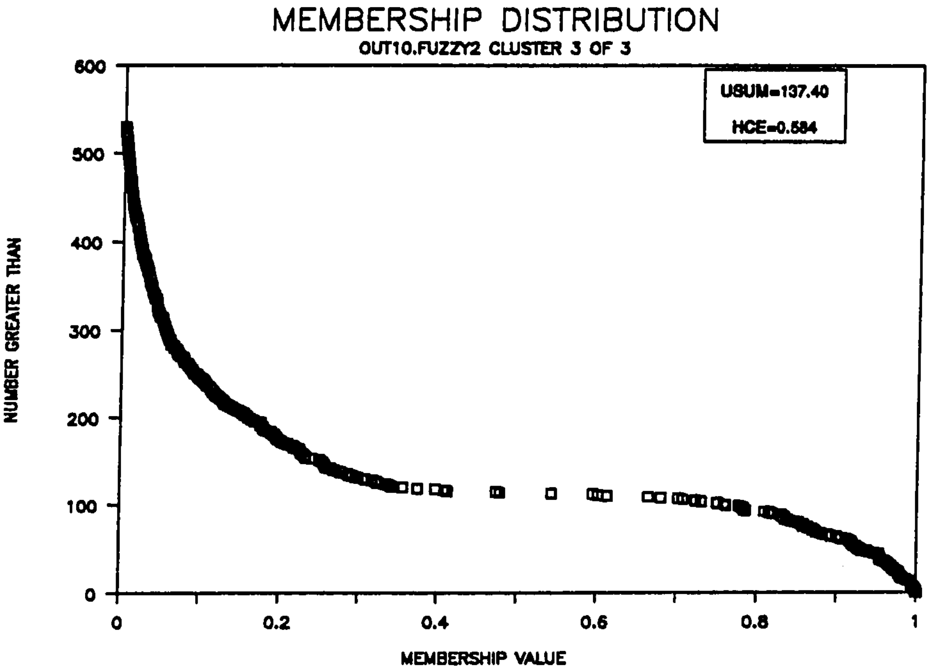
WALK	0.98	1.04	0.93
WAIT	0.69	0.85	0.55

PERCEPTION DIFFERENCES (ACTUAL - PERCEIVED):

WALK (min)	0.07	-0.16	0.27
WAIT (min)	1.80	1.09	3.17


MEMBERSHIP DISTRIBUTION
OUT10.FUZZY2 CLUSTER 1 OF 3
USUM=267.68
HCE=0.277
NUMBER GREATER THAN
0
100
200
300
400
500
600
0
0.2
0.4
0.6
0.8
1
MEMBERSHIP VALUE


MEMBERSHIP DISTRIBUTION
OUT10.FUZZY2 CLUSTER 2 OF 3
USUM=126.91
HCE=0.519
NUMBER GREATER THAN
0
100
200
300
400
500
600
0
0.2
0.4
0.6
0.8
1
MEMBERSHIP VALUE


MEMBERSHIP DISTRIBUTION
OUT10.FUZZY2 CLUSTER 3 OF 3
USUM=137.40
HCE=0.584
NUMBER GREATER THAN
600
500
400
300
200
100
0
0
0.2
0.4
0.6
0.8
1
MEMBERSHIP VALUE

FOUR CLUSTER SUMMARY

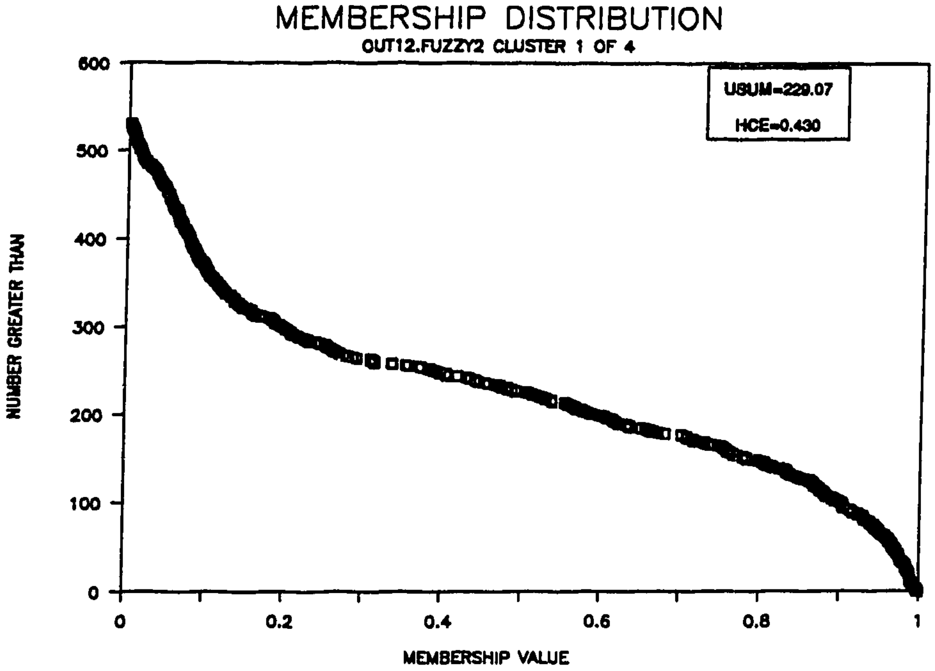
DIMENSION	CLUSTER			
	1	2	3	4
PWALK	3.309	3.506	4.209	3.536
PWAIT	3.959	4.097	5.659	5.059
AWALK	249.388	263.766	282.952	263.093
AWAIT	11.646	13.571	13.825	14.737
COMPU	-6.837	-7.533	-8.380	-8.118
INCOME	6.212	8.308	7.174	9.693
AGE	2774.436	3581.156	3291.513	3846.292
GENDER	1.953	1.203	1.716	1.339
JOB	4.952	4.930	4.901	3.398
MODE	0.961	0.900	0.170	0.517

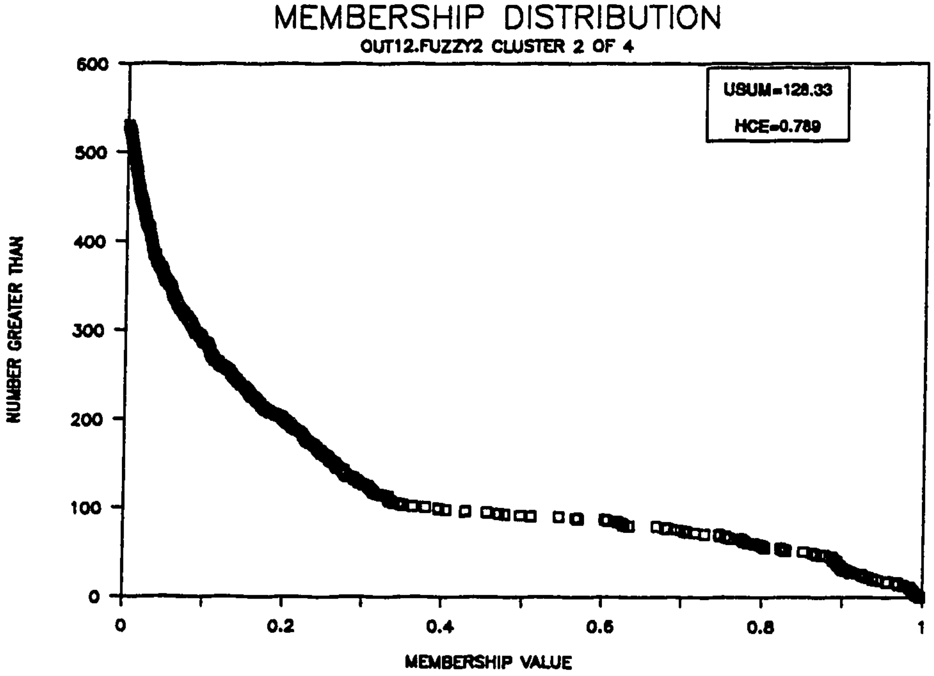
PERCEPTION RATIOS (PERCEIVED/ACTUAL):

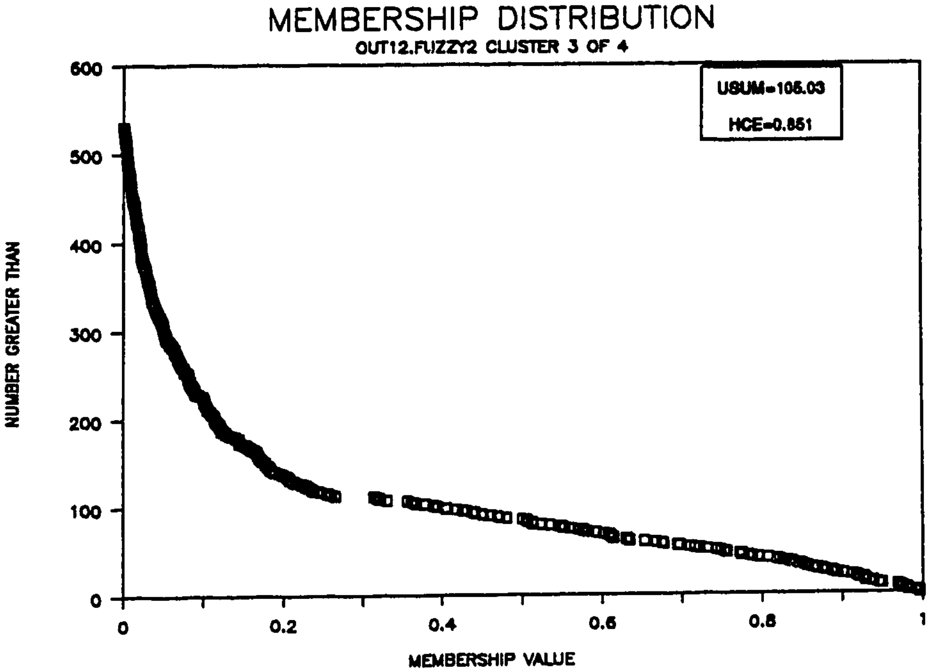
WALK	0.96	0.96	1.07	0.97
WAIT	0.68	0.60	0.82	0.69

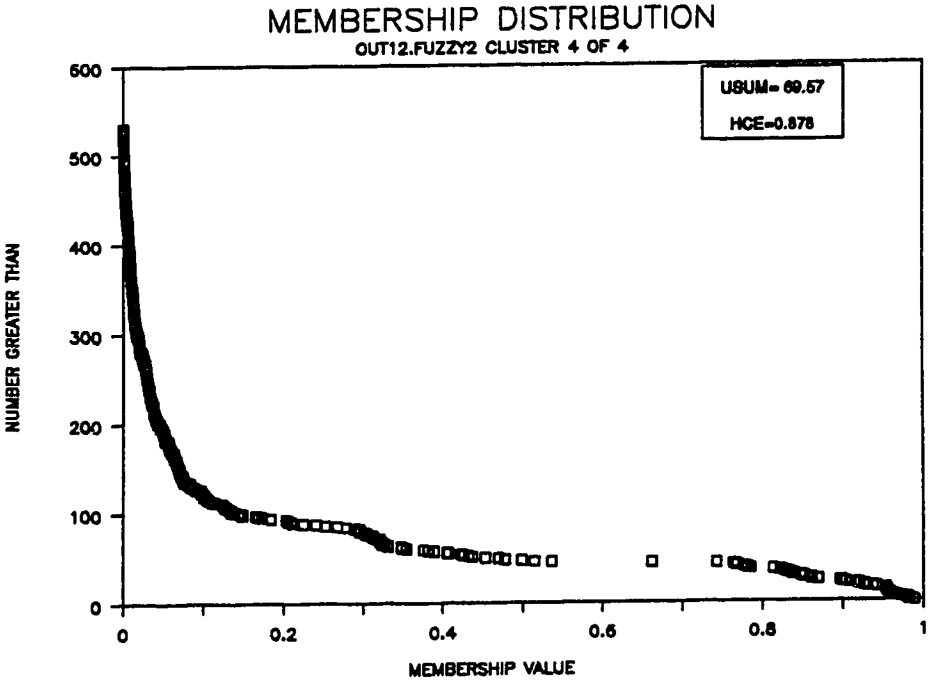
PERCEPTION DIFFERENCES (ACTUAL - PERCEIVED):

WALK (min)	0.15	0.16	-0.28	0.12
WAIT (min)	1.86	2.69	1.25	2.31


MEMBERSHIP DISTRIBUTION
OUT12.FUZZY2 CLUSTER 1 OF 4
USUM=229.07
HCE=0.430
NUMBER GREATER THAN
600
500
400
300
200
100
0
0
0.2
0.4
0.6
0.8
1
MEMBERSHIP VALUE


MEMBERSHIP DISTRIBUTION
OUT12.FUZZY2 CLUSTER 2 OF 4
USUM=128.33
HCE=0.789
NUMBER GREATER THAN
600
500
400
300
200
100
0
0
0.2
0.4
0.6
0.8
1
MEMBERSHIP VALUE


MEMBERSHIP DISTRIBUTION
OUT12.FUZZY2 CLUSTER 3 OF 4
USUM=105.03
HCE=0.851
NUMBER GREATER THAN
600
500
400
300
200
100
0
0
0.2
0.4
0.6
0.8
1
MEMBERSHIP VALUE


MEMBERSHIP DISTRIBUTION
OUT12.FUZZY2 CLUSTER 4 OF 4
USUM= 69.57
HCE=0.878
NUMBER GREATER THAN
600
500
400
300
200
100
0
0
0.2
0.4
0.6
0.8
1
MEMBERSHIP VALUE

FIVE CLUSTER SUMMARY

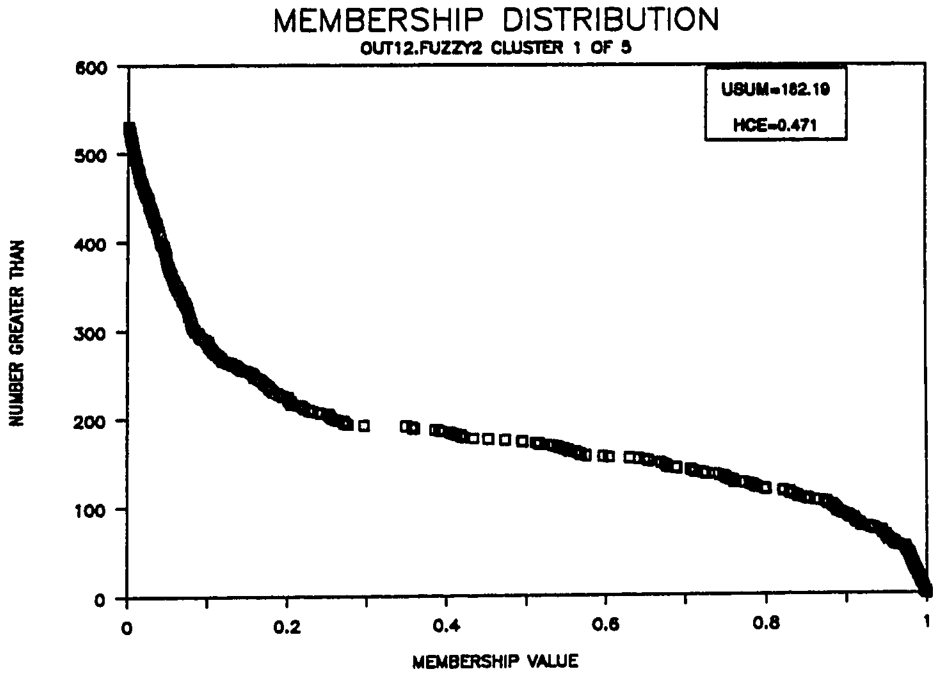
DIMENSION	CLUSTER				
	1	2	3	4	5
PWALK	3.135	4.118	4.113	3.306	3.417
PWAIT	3.755	4.907	5.639	3.931	4.938
AWALK	246.022	271.829	276.211	259.242	257.203
AWAIT	11.559	12.526	13.922	13.583	14.857
COMPU	-6.839	-6.956	-8.534	-7.592	-8.268
INCOME	6.189	6.257	7.361	8.753	10.164
AGE	2380.349	4467.285	3072.373	3372.977	3825.747
GENDER	1.955	1.893	1.680	1.090	1.262
JOB	4.948	4.897	4.889	4.929	3.217
MODE	0.962	0.905	0.090	0.903	0.460

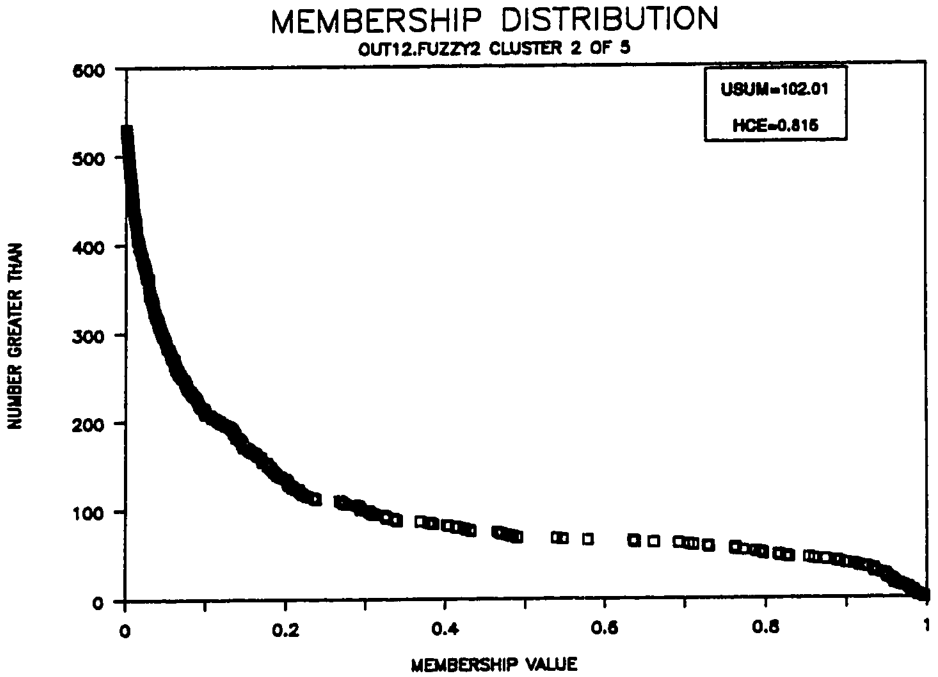
PERCEPTION RATIOS (PERCEIVED/ACTUAL):

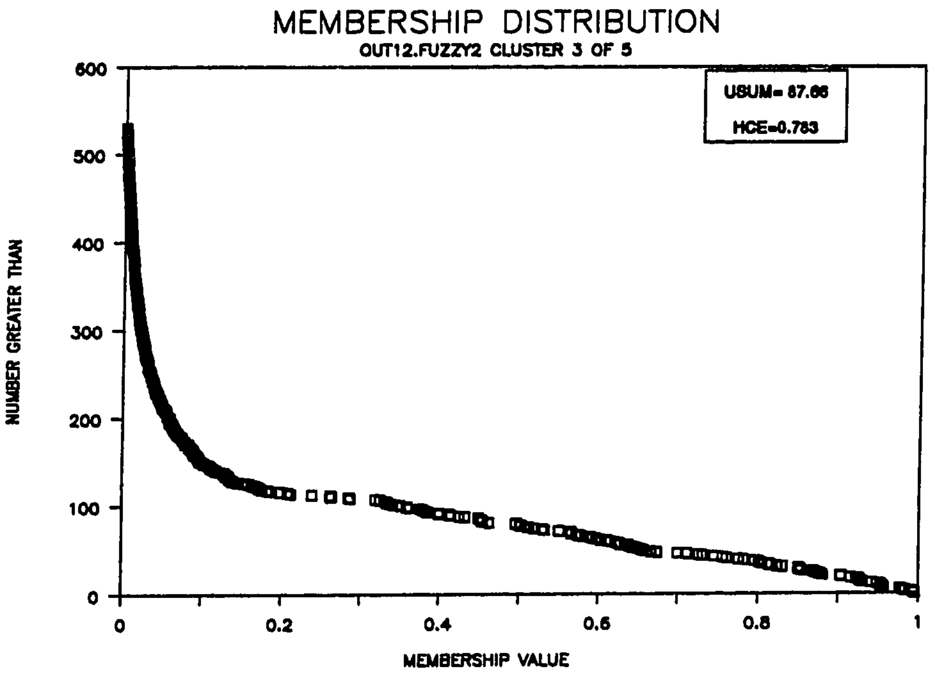
WALK	0.92	1.09	1.07	0.92	0.96
WAIT	0.65	0.78	0.81	0.58	0.67

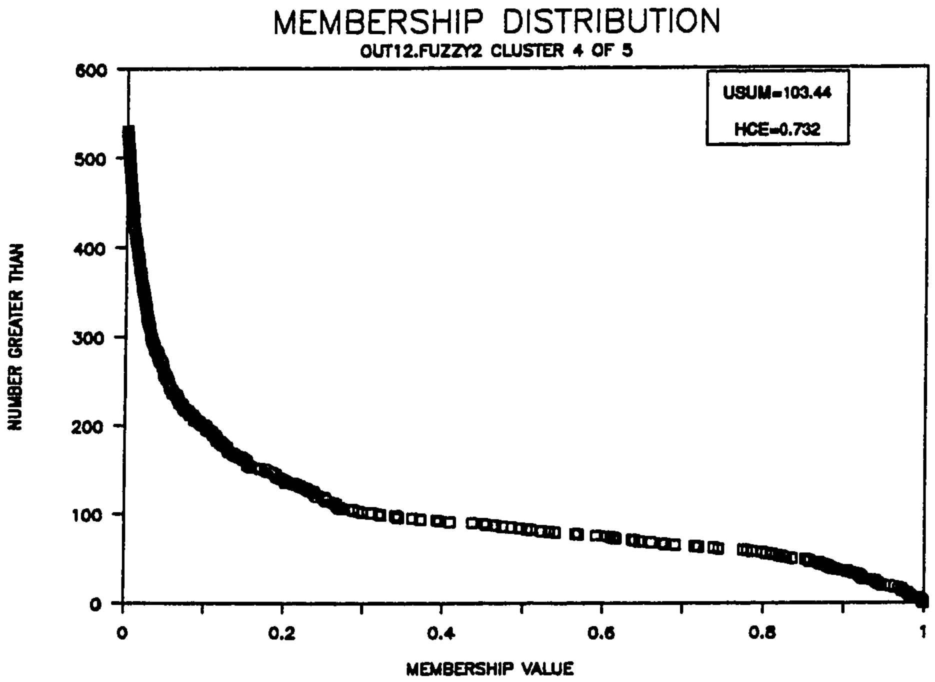
PERCEPTION DIFFERENCES (ACTUAL - PERCEIVED):

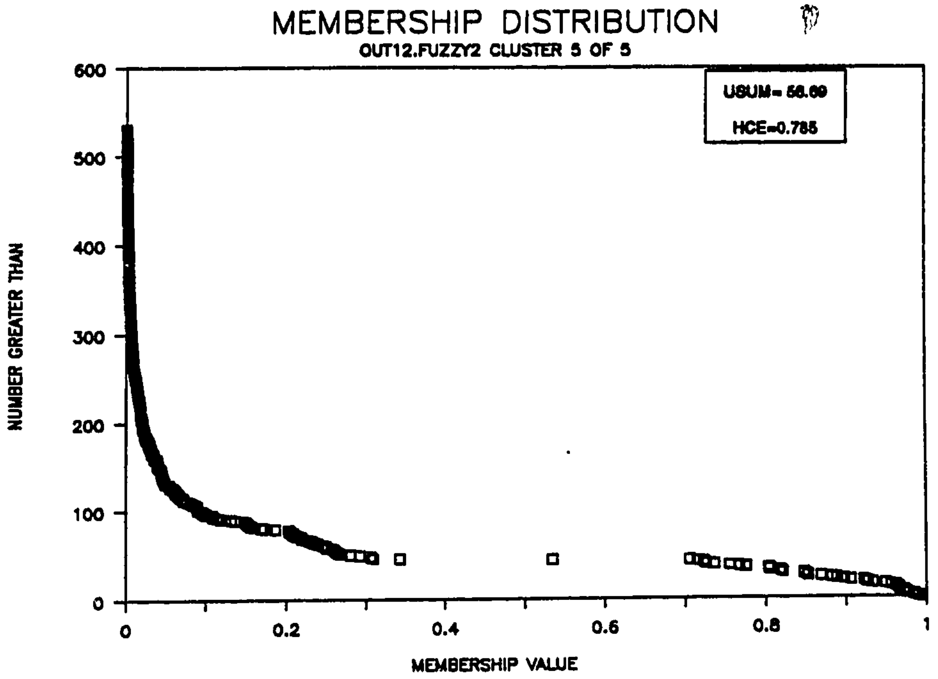
WALK (min)	0.28	-0.34	-0.28	0.29	0.16
WAIT (min)	2.02	1.36	1.32	2.86	2.49


MEMBERSHIP DISTRIBUTION
OUT12.FUZZY2 CLUSTER 1 OF 5
USUM=182.19
HCE=0.471
NUMBER GREATER THAN
600
500
400
300
200
100
0
0
0.2
0.4
0.6
0.8
1
MEMBERSHIP VALUE


MEMBERSHIP DISTRIBUTION
OUT12.FUZZY2 CLUSTER 2 OF 5
USUM=102.01
HCE=0.815
NUMBER GREATER THAN
600
500
400
300
200
100
0
0
0.2
0.4
0.6
0.8
1
MEMBERSHIP VALUE


MEMBERSHIP DISTRIBUTION
OUT12.FUZZY2 CLUSTER 3 OF 5
USUM= 87.66
HCE=0.783
NUMBER GREATER THAN
600
500
400
300
200
100
0
0
0.2
0.4
0.6
0.8
1
MEMBERSHIP VALUE


MEMBERSHIP DISTRIBUTION
OUT12.FUZZY2 CLUSTER 4 OF 5
USUM=103.44
HCE=0.732
NUMBER GREATER THAN
600
500
400
300
200
100
0
0
0.2
0.4
0.6
0.8
1
MEMBERSHIP VALUE


MEMBERSHIP DISTRIBUTION
OUT12.FUZZY2 CLUSTER 5 OF 5
USUM= 56.69
HCE=0.785
NUMBER GREATER THAN
600
500
400
300
200
100
0
0
0.2
0.4
0.6
0.8
1
MEMBERSHIP VALUE

SIX CLUSTER SUMMARY

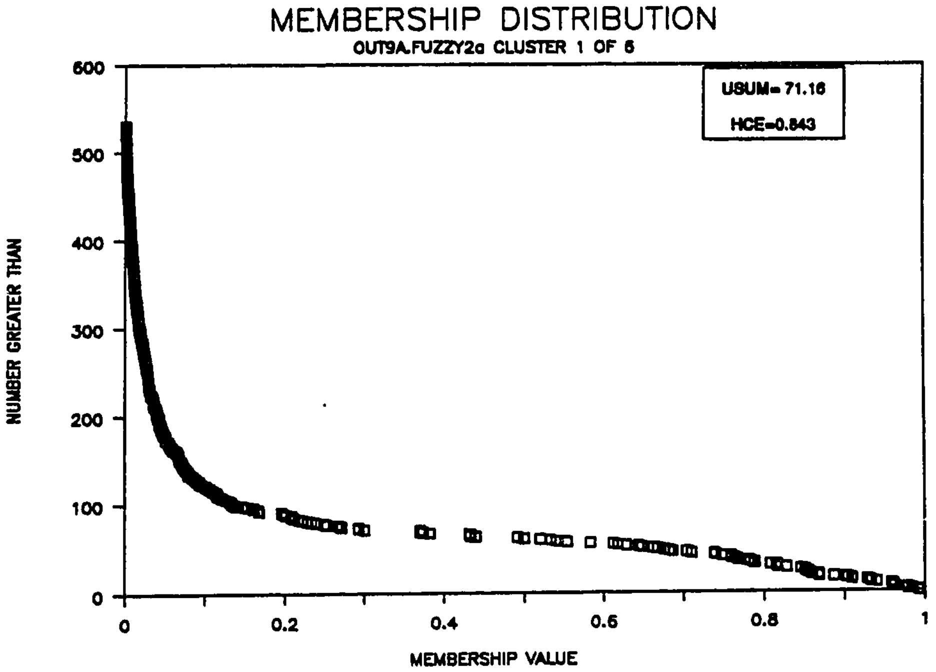
DIMENSION	CLUSTER					
	1	2	3	4	5	6
PWALK	3.424	3.965	3.992	3.443	3.391	3.187
PWAIT	4.380	4.750	5.691	3.896	4.866	3.748
AWALK	243.819	275.327	265.556	264.552	256.469	261.607
AWAIT	12.441	12.847	13.781	13.669	14.761	11.348
COMPU	-6.953	-6.941	-8.445	-7.641	-8.261	-6.971
INCOME	6.661	6.183	7.361	8.933	10.238	6.193
AGE	2666.496	4583.543	3083.148	3431.908	3853.556	2399.054
GENDER	1.797	1.886	1.715	1.087	1.242	1.941
JOB	4.921	4.896	4.895	4.922	3.187	4.941
MODE	0.876	0.904	0.083	0.893	0.444	0.957
MODEFREQ	2.611	3.92	3.917	3.847	3.842	3.948

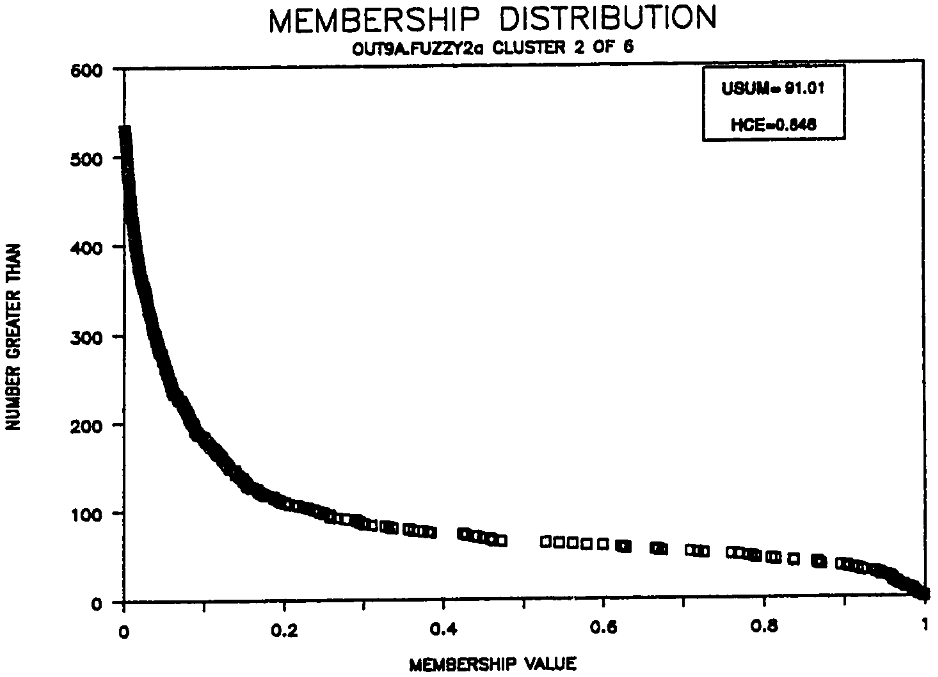
PERCEPTION RATIOS (PERCEIVED/ACTUAL):

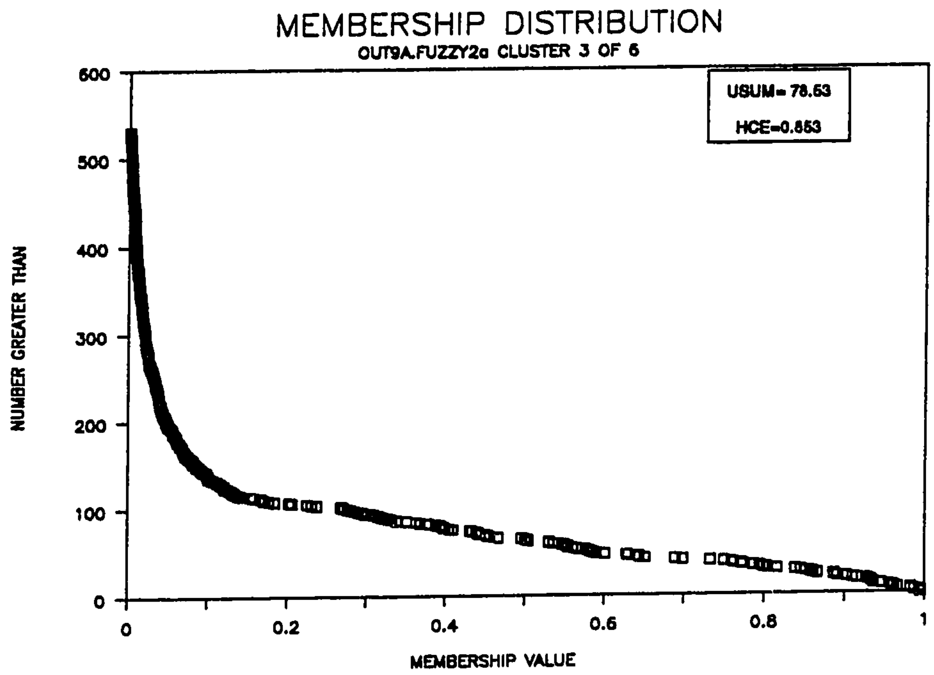
WALK	1.01	1.04	1.08	0.94	0.95	0.88
WAIT	0.70	0.74	0.83	0.57	0.66	0.66

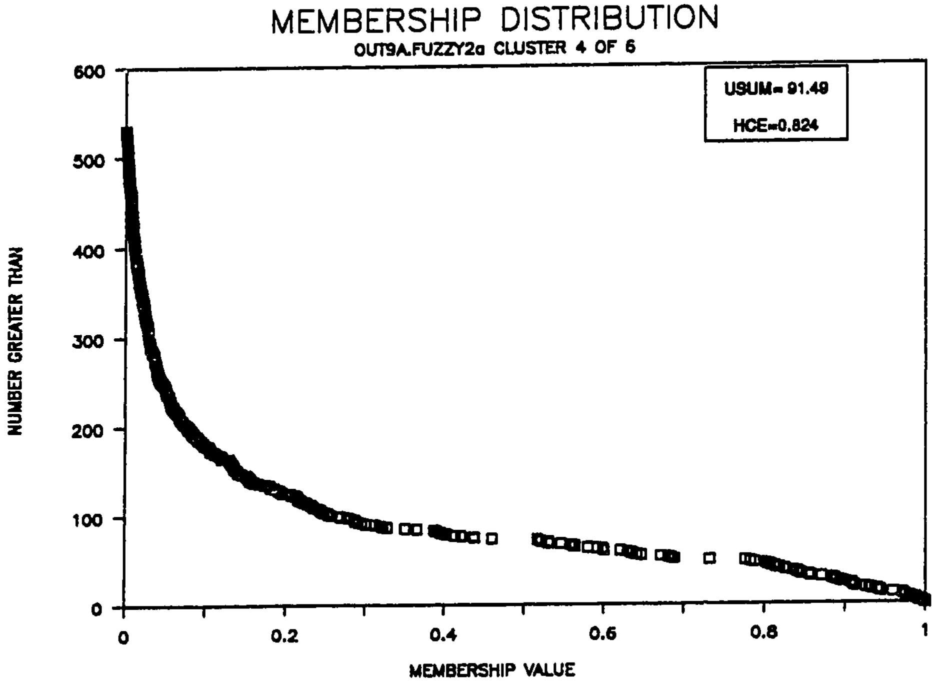
PERCEPTION DIFFERENCES (ACTUAL - PERCEIVED):

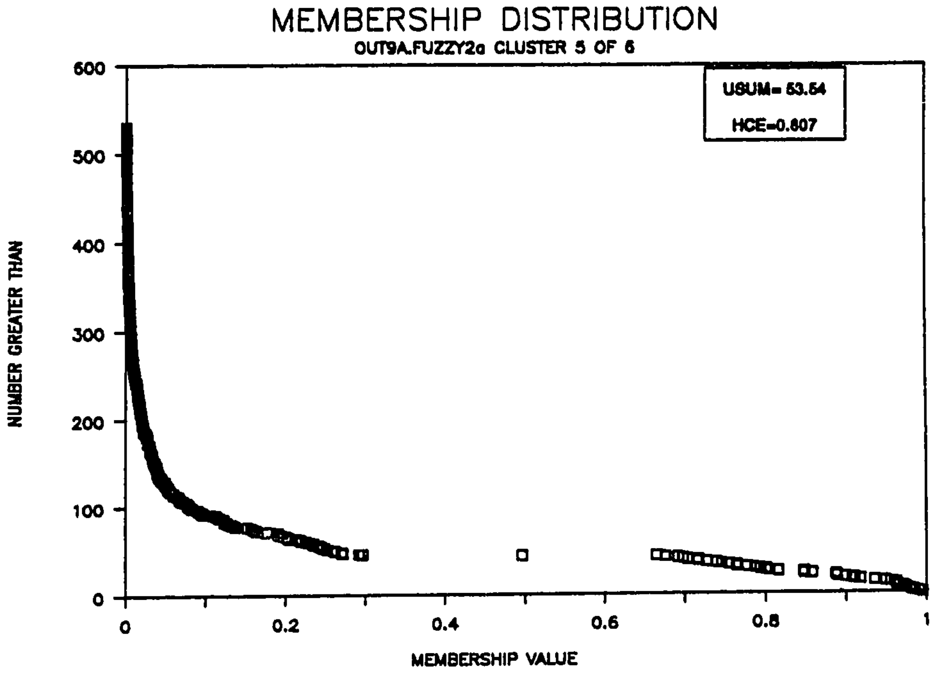
WALK (min)	-0.04	-0.14	-0.30	0.23	0.17	0.45
WAIT (min)	1.84	1.67	1.20	2.94	2.51	1.93

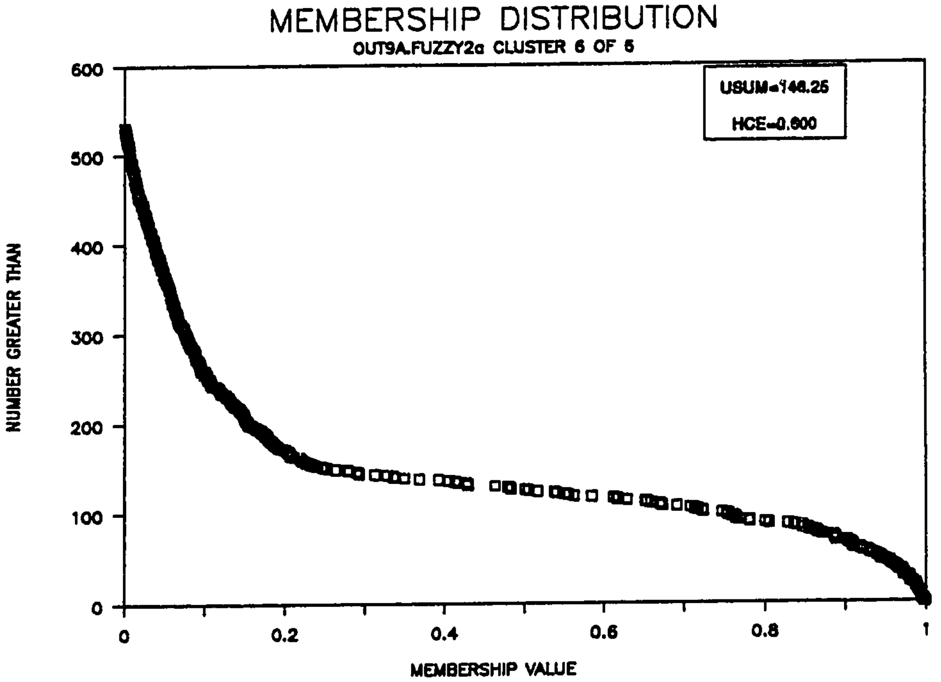

MEMBERSHIP DISTRIBUTION
OUT9A.FUZZY2a CLUSTER 1 OF 6
USUM= 71.16
HCE=0.843
NUMBER GREATER THAN
600
500
400
300
200
100
0
0
0.2
0.4
0.6
0.8
1
MEMBERSHIP VALUE


MEMBERSHIP DISTRIBUTION
OUT9A.FUZZY2a CLUSTER 2 OF 6
USUM= 91.01
HCE=0.846
NUMBER GREATER THAN
600
500
400
300
200
100
0
0
0.2
0.4
0.6
0.8
1
MEMBERSHIP VALUE

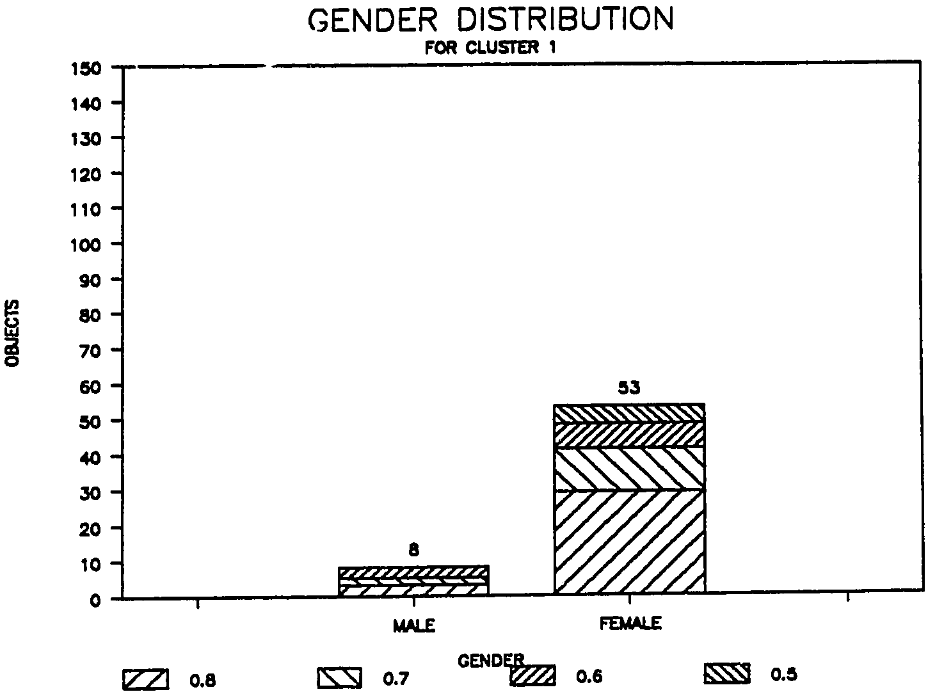

MEMBERSHIP DISTRIBUTION
OUT9A.FUZZY2a CLUSTER 3 OF 6
USUM= 78.63
HCE=0.853
NUMBER GREATER THAN
600
500
400
300
200
100
0
0
0.2
0.4
0.6
0.8
1
MEMBERSHIP VALUE

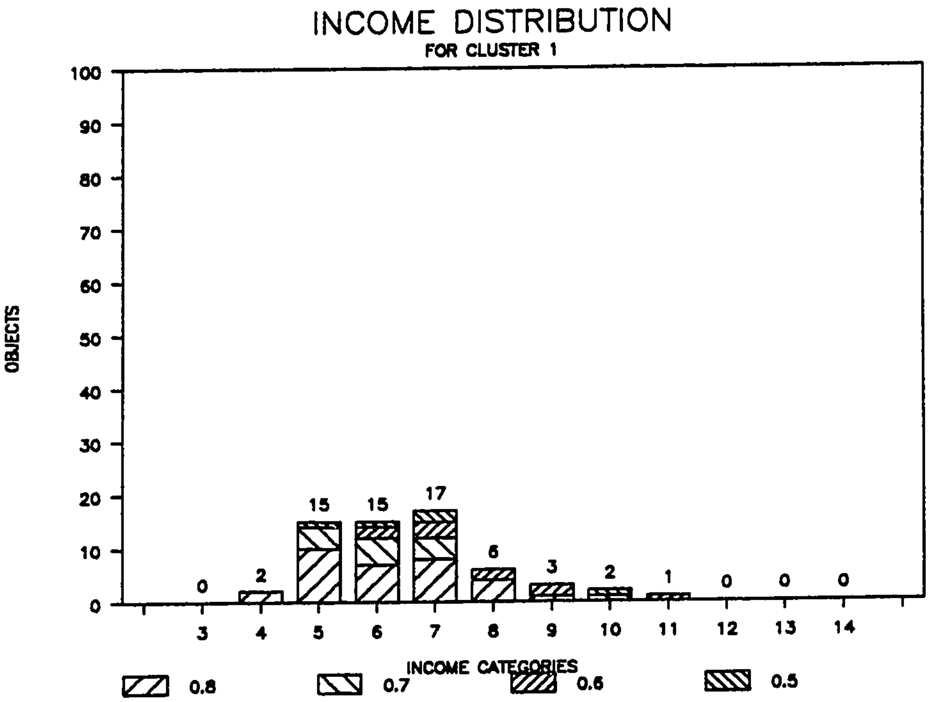

MEMBERSHIP DISTRIBUTION
OUT9A.FUZZY2a CLUSTER 4 OF 6
USUM= 91.49
HCE=0.824
NUMBER GREATER THAN
600
500
400
300
200
100
0
0
0.2
0.4
0.6
0.8
1
MEMBERSHIP VALUE

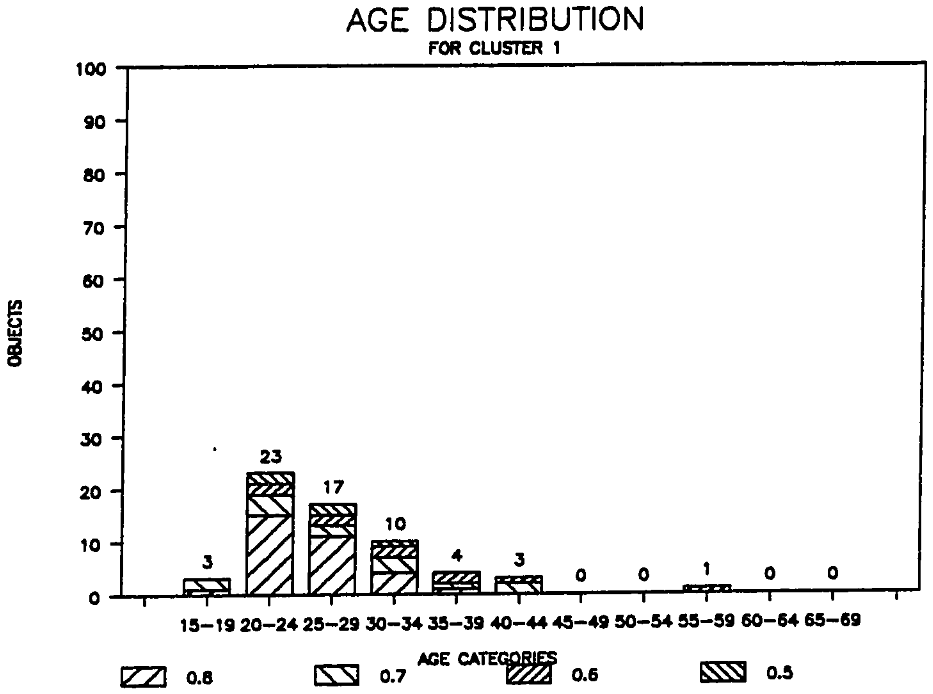

MEMBERSHIP DISTRIBUTION
OUT9A.FUZZY2a CLUSTER 5 OF 6
USUM= 53.54
HCE=0.607
NUMBER GREATER THAN
600
500
400
300
200
100
0
MEMBERSHIP VALUE
0
0.2
0.4
0.6
0.8
1

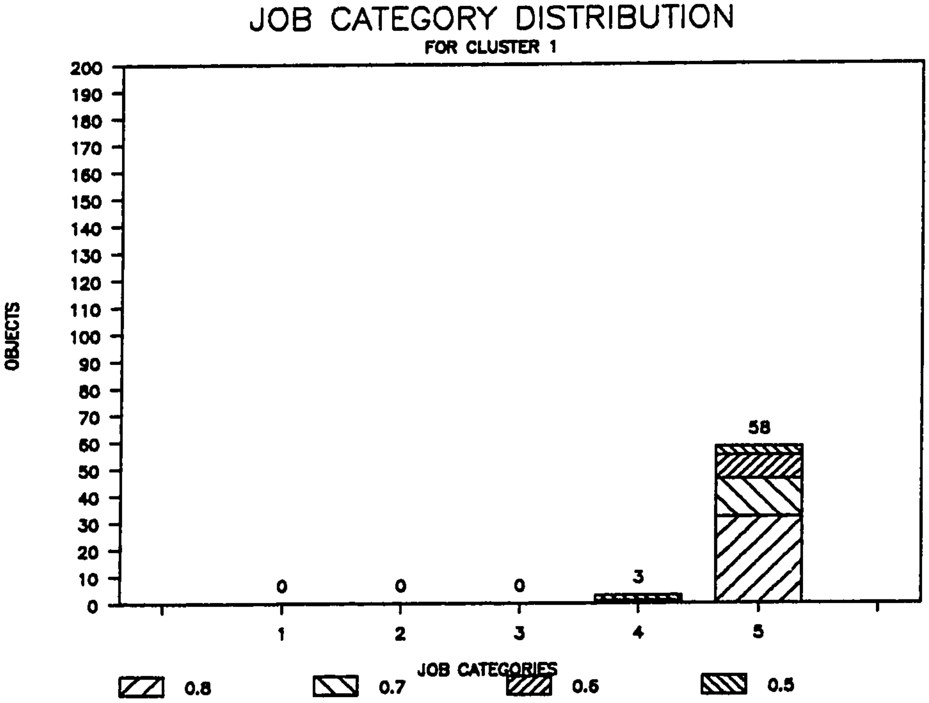

MEMBERSHIP DISTRIBUTION
OUT9A.FUZZY2a CLUSTER 6 OF 6
USUM=146.25
HCE=0.600
NUMBER GREATER THAN
600
500
400
300
200
100
0
MEMBERSHIP VALUE
0
0.2
0.4
0.6
0.8
1

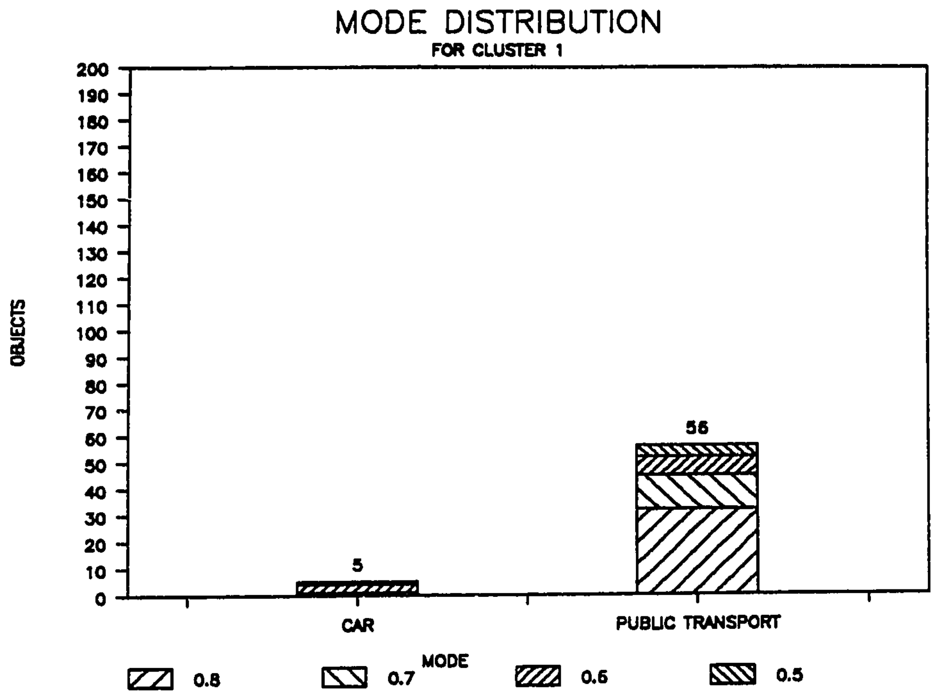
Appendix C
Socioeconomic and Mode Choice Distributions for the Six Cluster Solution Cluster Cores

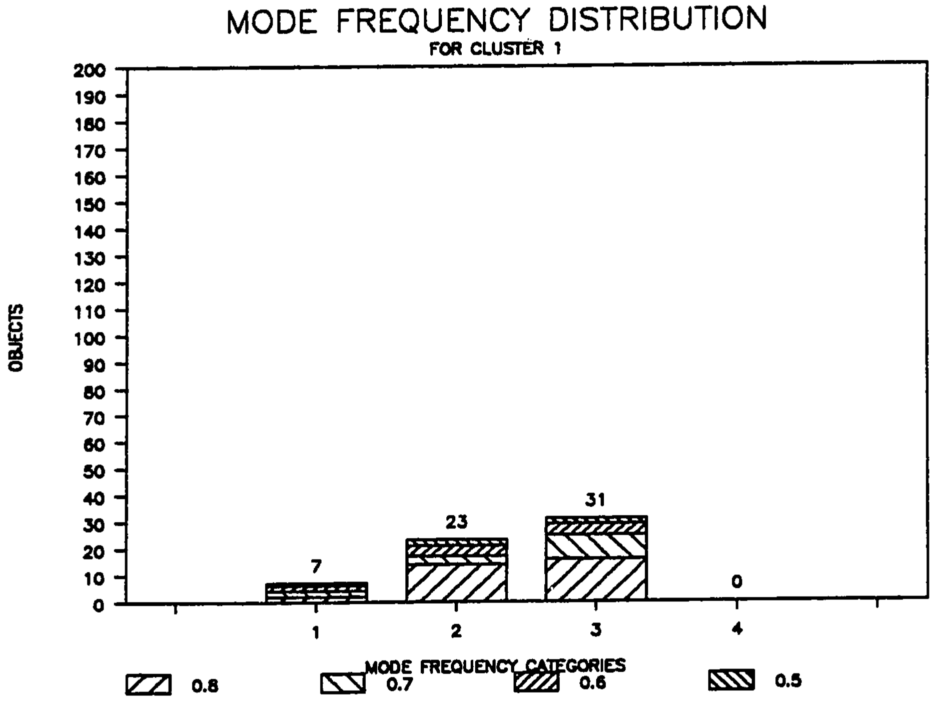

GENDER DISTRIBUTION
FOR CLUSTER 1
OBJECTS
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
8
53
MALE
FEMALE
GENDER
0.8
0.7
0.6
0.5

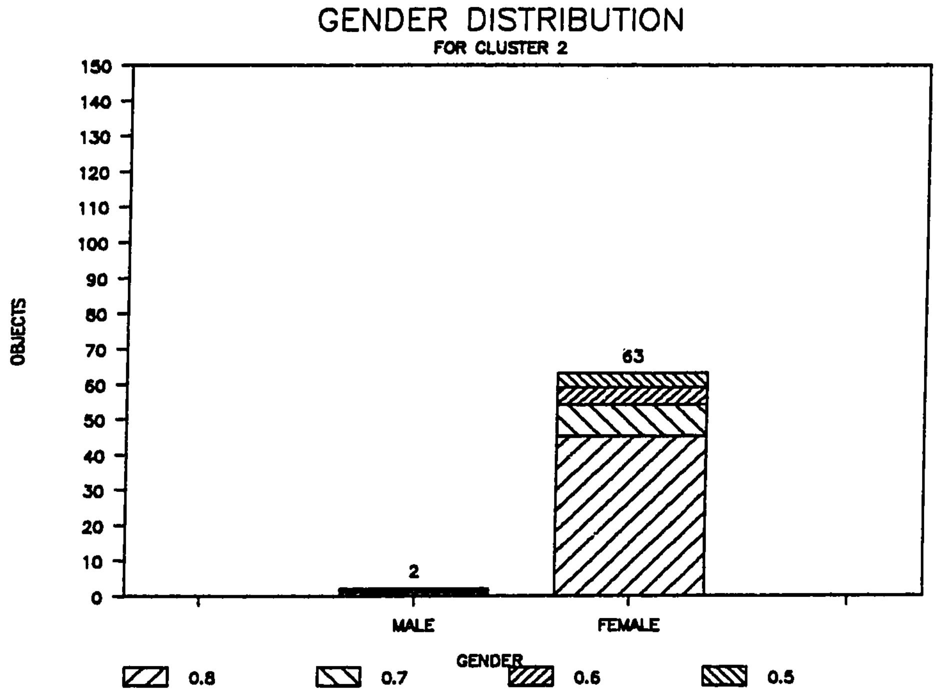

INCOME DISTRIBUTION
FOR CLUSTER 1
OBJECTS
100
90
80
70
60
50
40
30
20
10
0
0
2
15
15
17
6
3
2
1
0
0
0
3
4
5
6
7
8
9
10
11
12
13
14
INCOME CATEGORIES
0.8
0.7
0.6
0.5

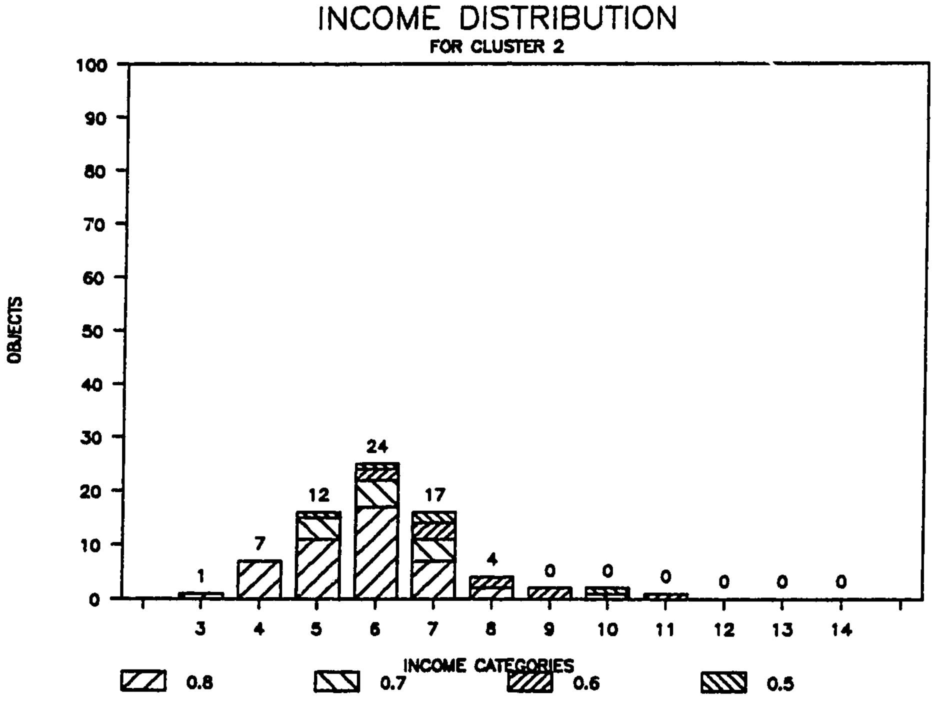

AGE DISTRIBUTION
FOR CLUSTER 1
OBJECTS
100
90
80
70
60
50
40
30
20
10
0
3
23
17
10
4
3
0
0
1
0
0
15–19
20–24
25–29
30–34
35–39
40–44
45–49
50–54
55–59
60–64
65–69
AGE CATEGORIES
0.8
0.7
0.6
0.5

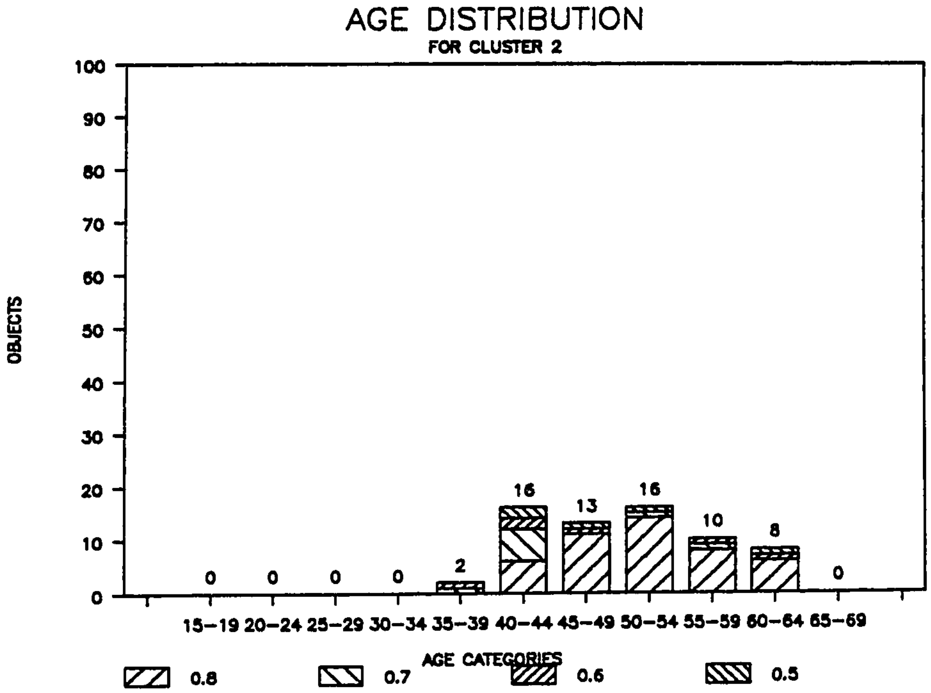

JOB CATEGORY DISTRIBUTION
FOR CLUSTER 1
OBJECTS
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
0
0
0
3
58
1
2
3
4
5
JOB CATEGORIES
0.8
0.7
0.6
0.5

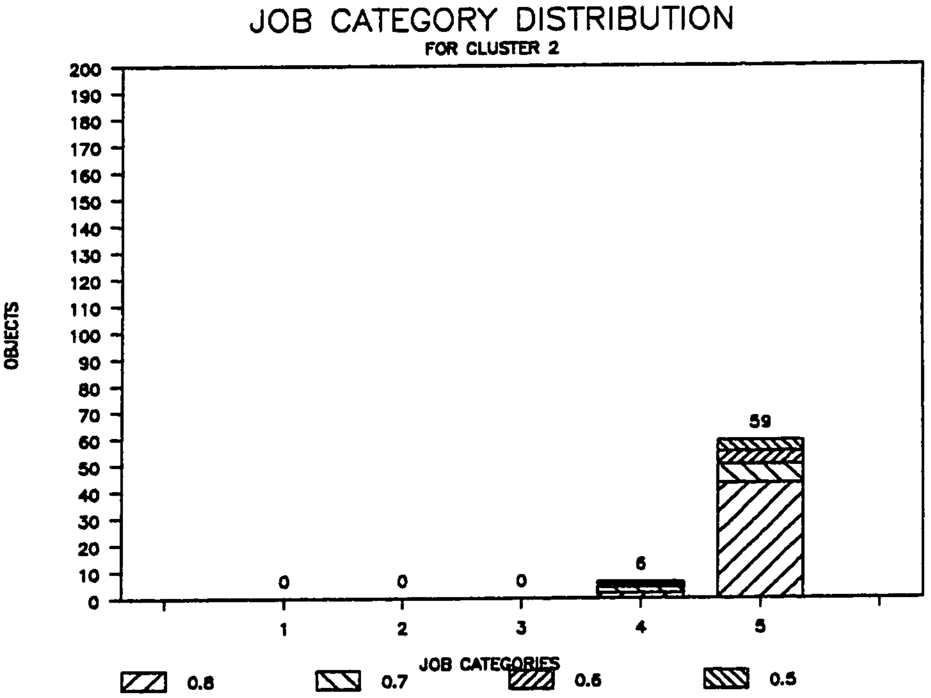

MODE DISTRIBUTION
FOR CLUSTER 1
OBJECTS
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
5
56
CAR
PUBLIC TRANSPORT
MODE
0.8
0.7
0.6
0.5

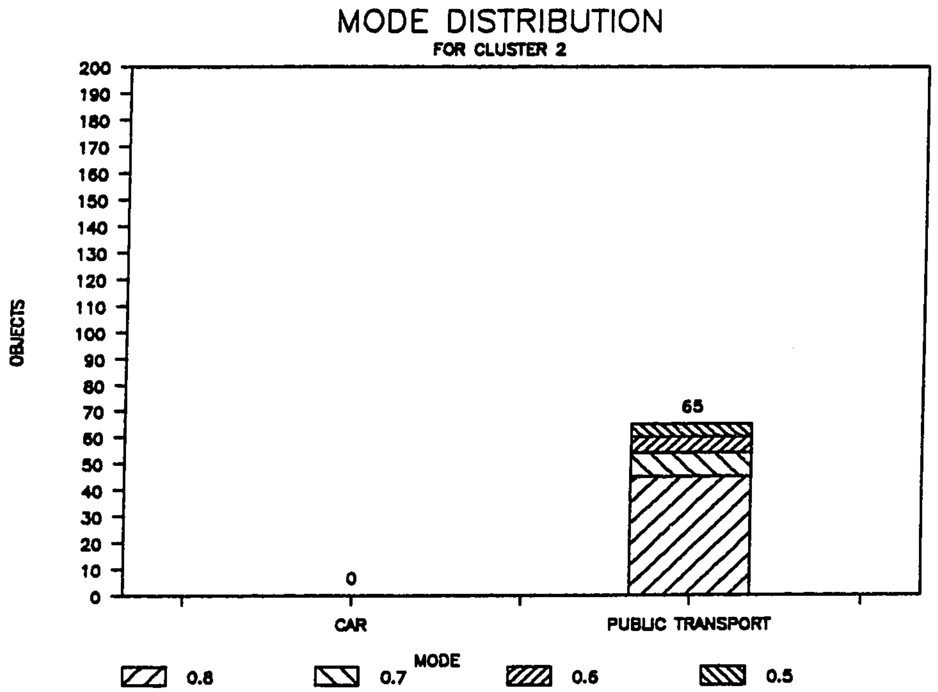

MODE FREQUENCY DISTRIBUTION
FOR CLUSTER 1
OBJECTS
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
7
23
31
0
1
2
3
4
MODE FREQUENCY CATEGORIES
0.8
0.7
0.6
0.5

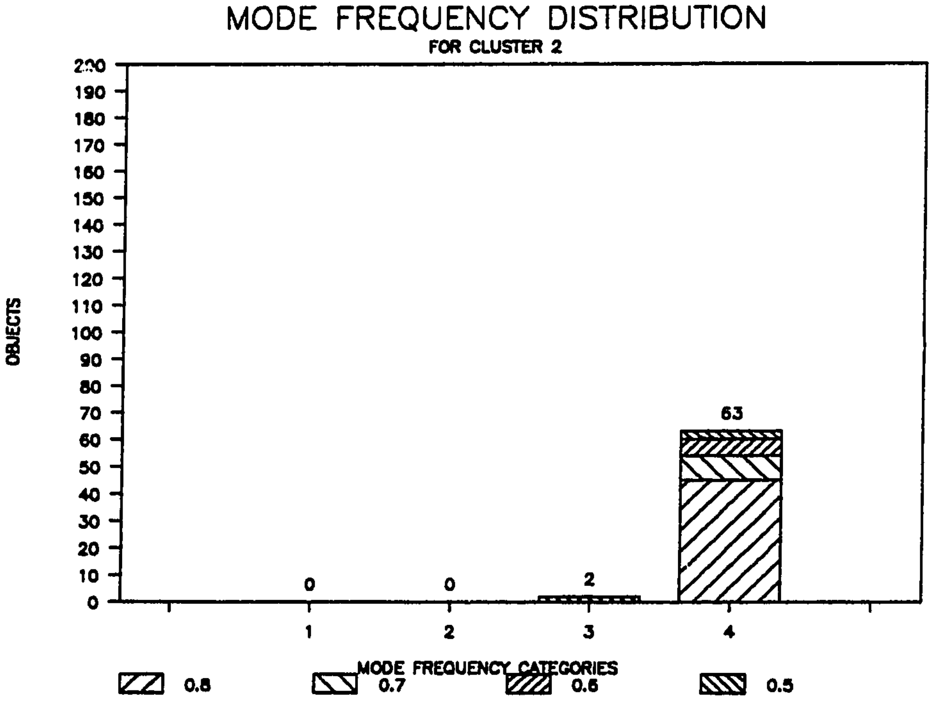

GENDER DISTRIBUTION
FOR CLUSTER 2
OBJECTS
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
2
63
MALE
FEMALE
GENDER
0.8
0.7
0.6
0.5

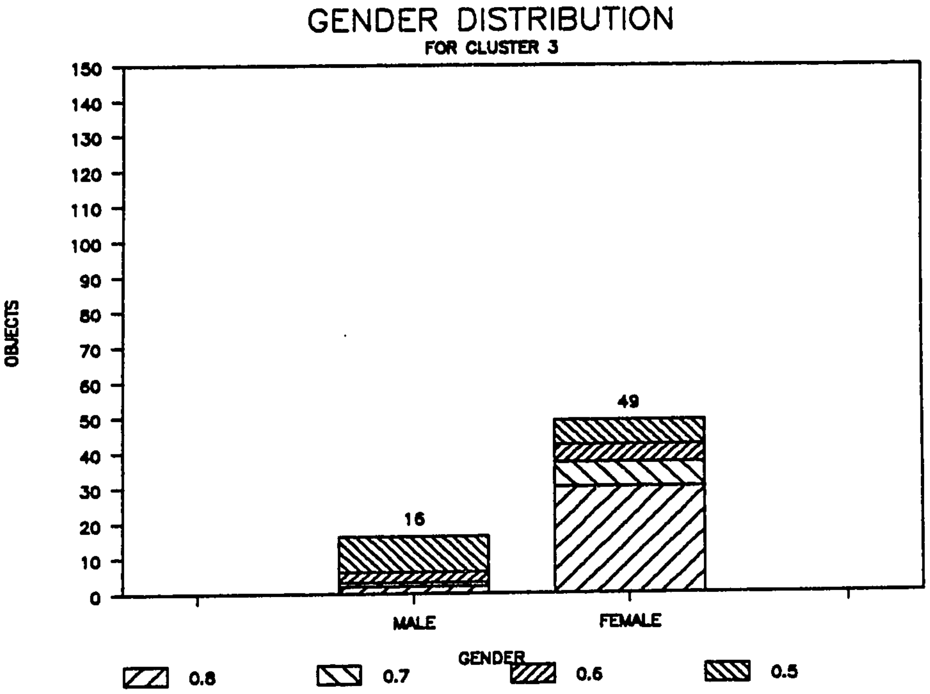

INCOME DISTRIBUTION
FOR CLUSTER 2
OBJECTS
100
90
80
70
60
50
40
30
20
10
0
1
7
12
24
17
4
0
0
0
0
0
0
3
4
5
6
7
8
9
10
11
12
13
14
INCOME CATEGORIES
0.8
0.7
0.6
0.5

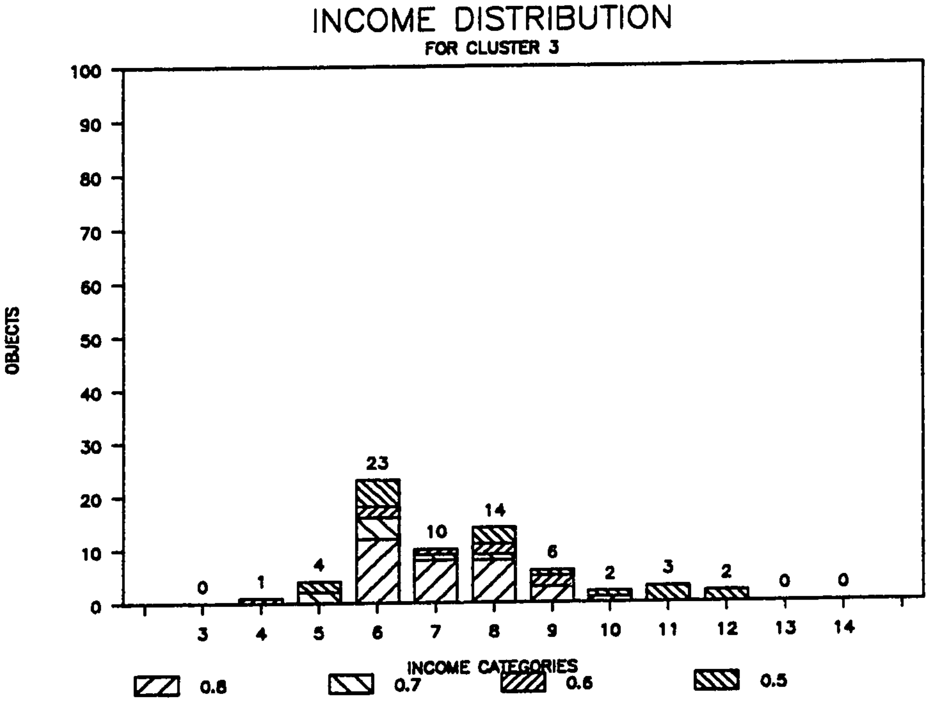

AGE DISTRIBUTION
FOR CLUSTER 2
OBJECTS
100
90
80
70
60
50
40
30
20
10
0
0
0
0
0
2
16
13
16
10
8
0
15–19
20–24
25–29
30–34
35–39
40–44
45–49
50–54
55–59
60–64
65–69
AGE CATEGORIES
0.8
0.7
0.6
0.5


JOB CATEGORY DISTRIBUTION
FOR CLUSTER 2
OBJECTS
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
0
0
0
6
59
1
2
3
4
5
JOB CATEGORIES
0.8
0.7
0.6
0.5

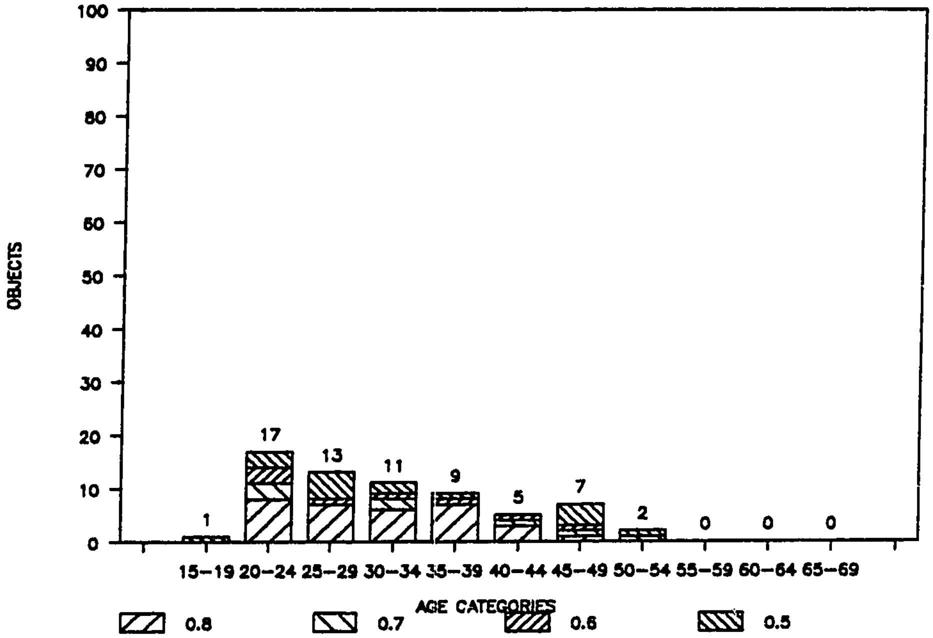

MODE DISTRIBUTION
FOR CLUSTER 2
OBJECTS
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
0
65
CAR
PUBLIC TRANSPORT
MODE
▨ 0.8
▧ 0.7
▨ 0.6
▧ 0.5


MODE FREQUENCY DISTRIBUTION
FOR CLUSTER 2
OBJECTS
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
0
0
2
63
1
2
3
4
MODE FREQUENCY CATEGORIES
▨ 0.8
▧ 0.7
▨ 0.6
▧ 0.5

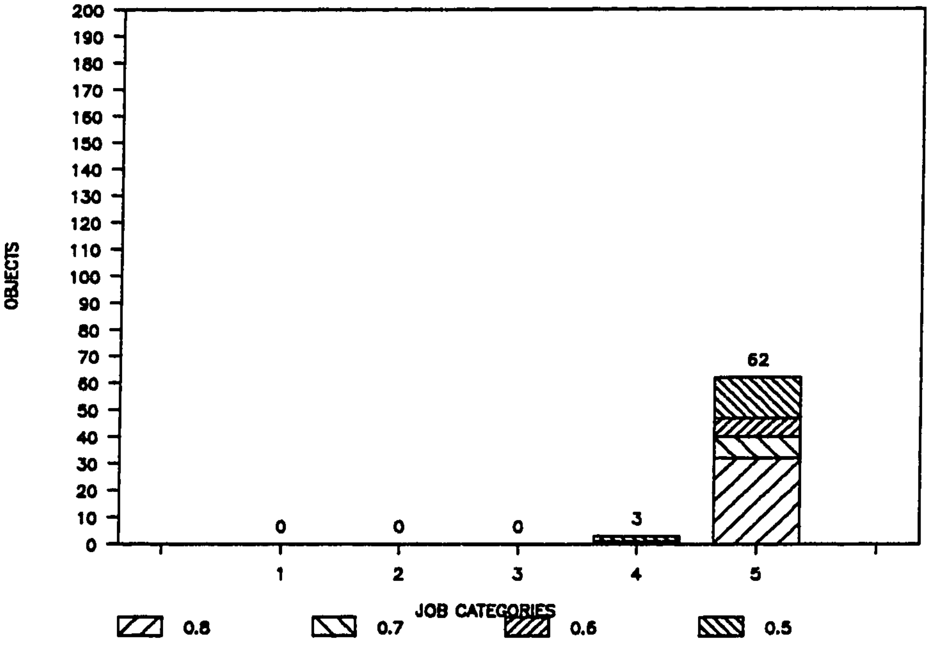

GENDER DISTRIBUTION
FOR CLUSTER 3
OBJECTS
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
16
49
MALE
FEMALE
GENDER
0.8
0.7
0.6
0.5

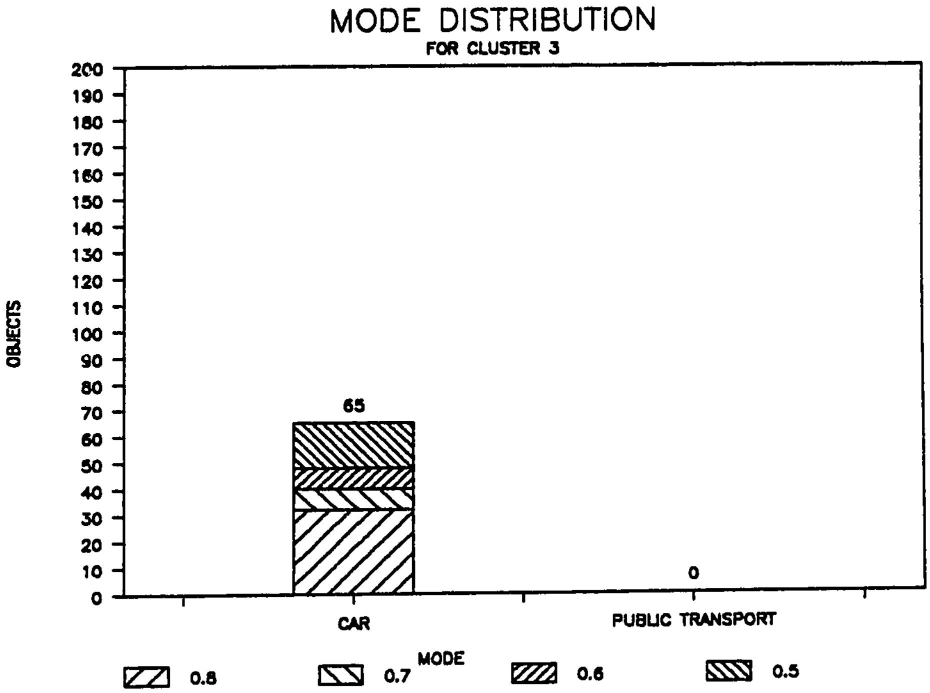

INCOME DISTRIBUTION
FOR CLUSTER 3
OBJECTS
0
10
20
30
40
50
60
70
80
90
100
0
1
4
23
10
14
6
2
3
2
0
0
3
4
5
6
7
8
9
10
11
12
13
14
INCOME CATEGORIES
0.8
0.7
0.6
0.5

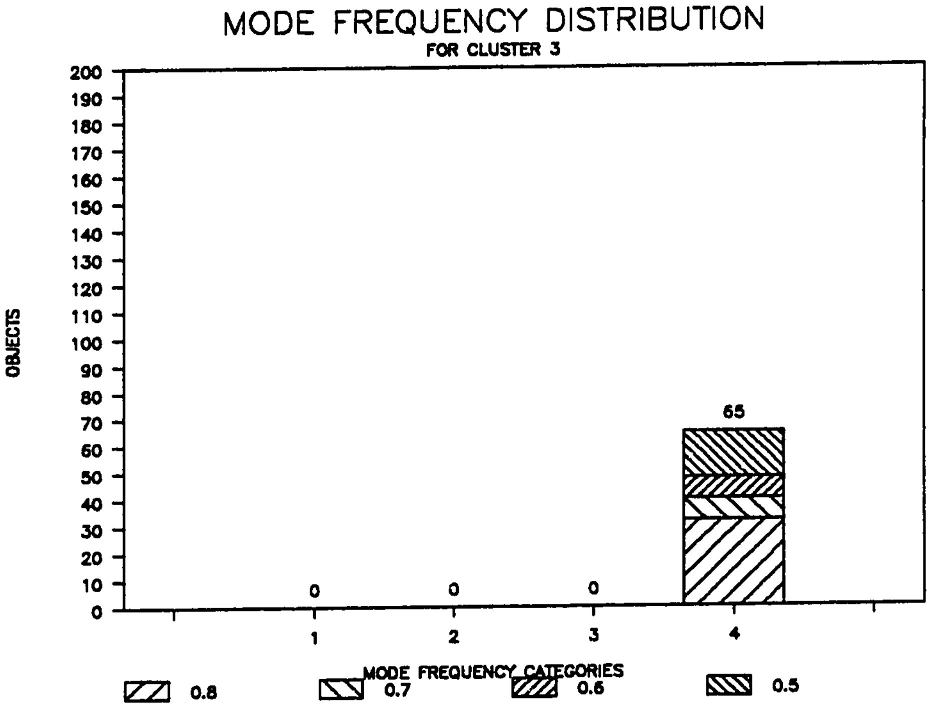
AGE DISTRIBUTION
FOR CLUSTER 3

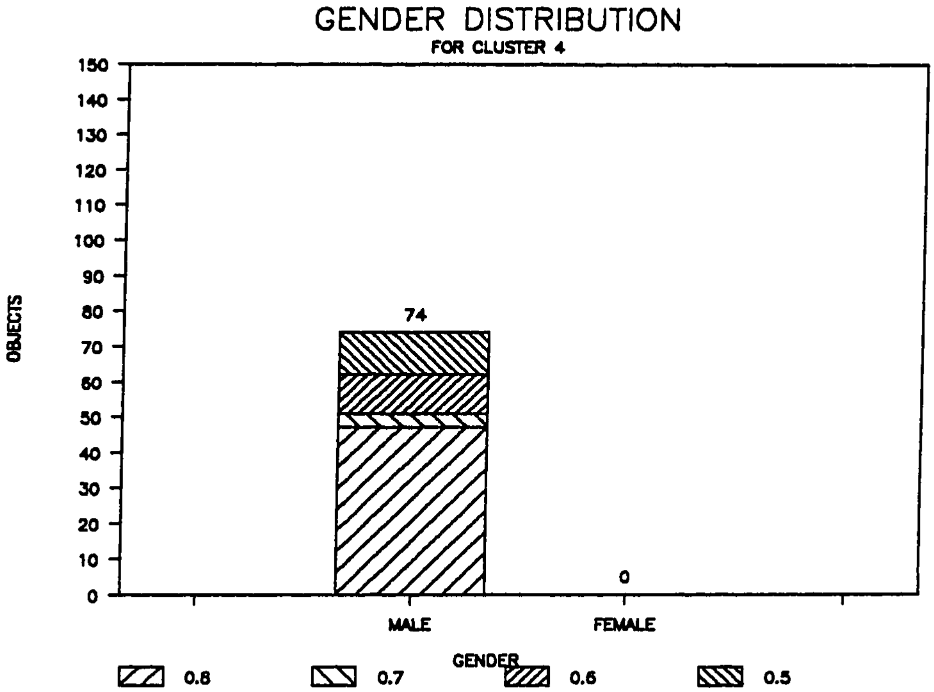


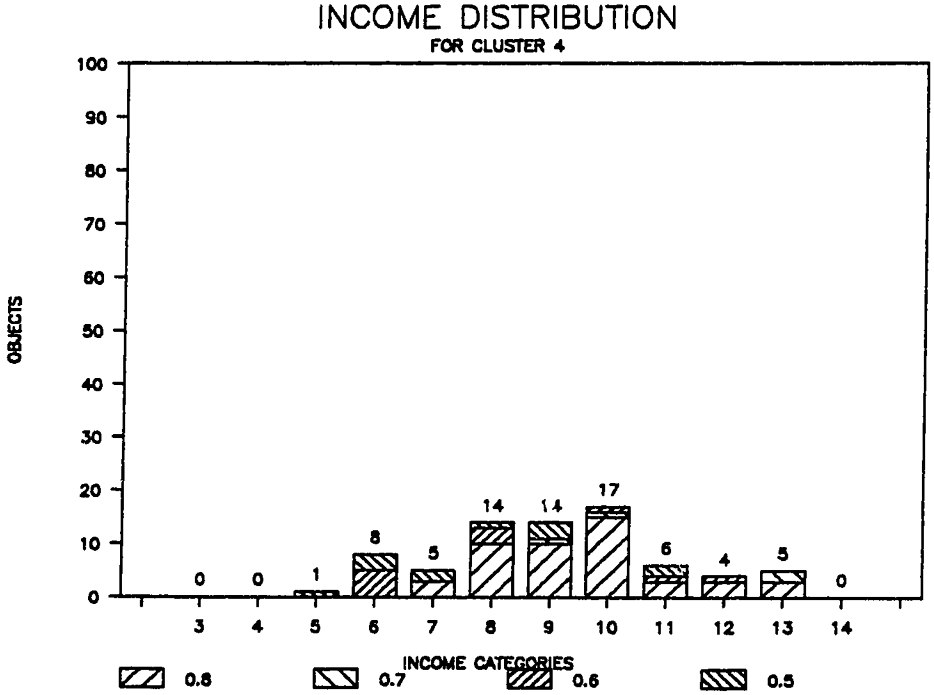
JOB CATEGORY DISTRIBUTION
FOR CLUSTER 3

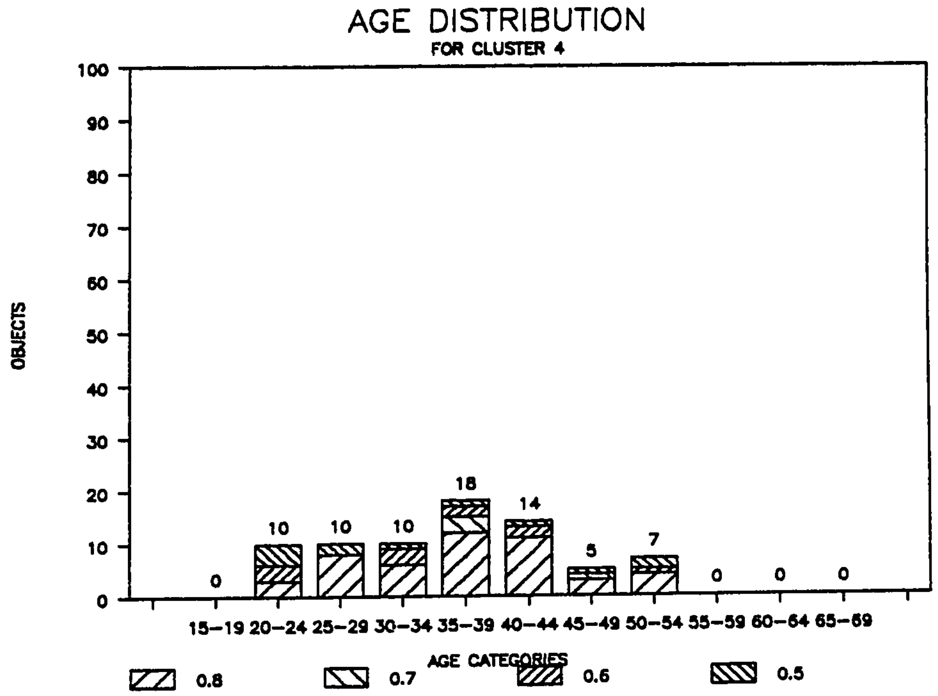


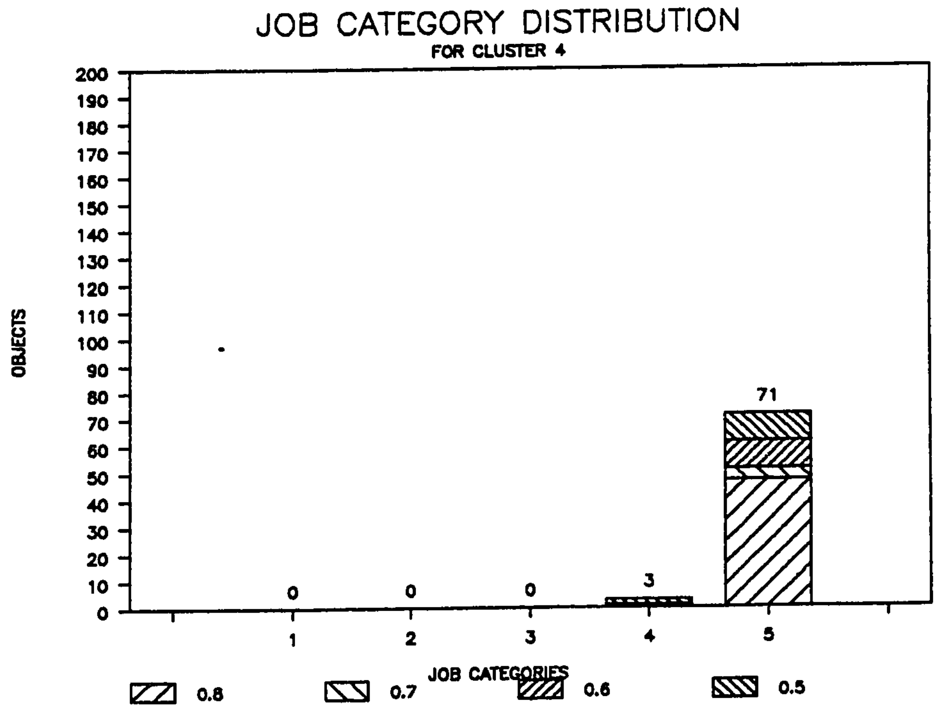

MODE DISTRIBUTION
FOR CLUSTER 3
OBJECTS
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
65
0
CAR
PUBLIC TRANSPORT
MODE
0.8
0.7
0.6
0.5

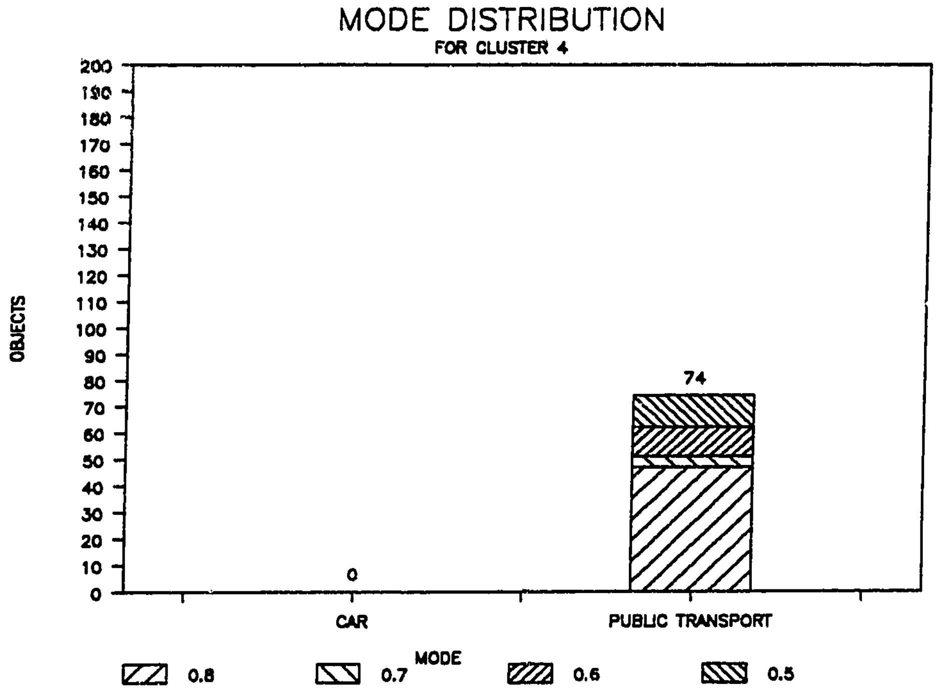

MODE FREQUENCY DISTRIBUTION
FOR CLUSTER 3
OBJECTS
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
0
0
0
65
1
2
3
4
MODE FREQUENCY CATEGORIES
0.8
0.7
0.6
0.5

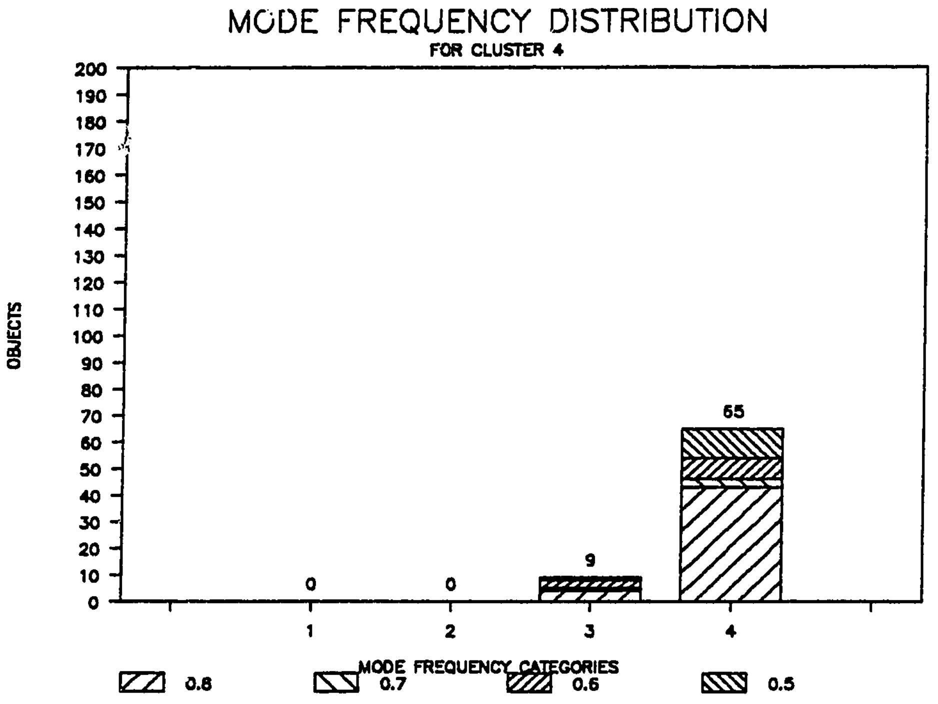

GENDER DISTRIBUTION
FOR CLUSTER 4
OBJECTS
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
74
0
MALE
FEMALE
GENDER
0.8
0.7
0.6
0.5


INCOME DISTRIBUTION
FOR CLUSTER 4
OBJECTS
0
10
20
30
40
50
60
70
80
90
100
0
0
1
8
5
14
14
17
6
4
5
0
3
4
5
6
7
8
9
10
11
12
13
14
INCOME CATEGORIES
0.8
0.7
0.6
0.5

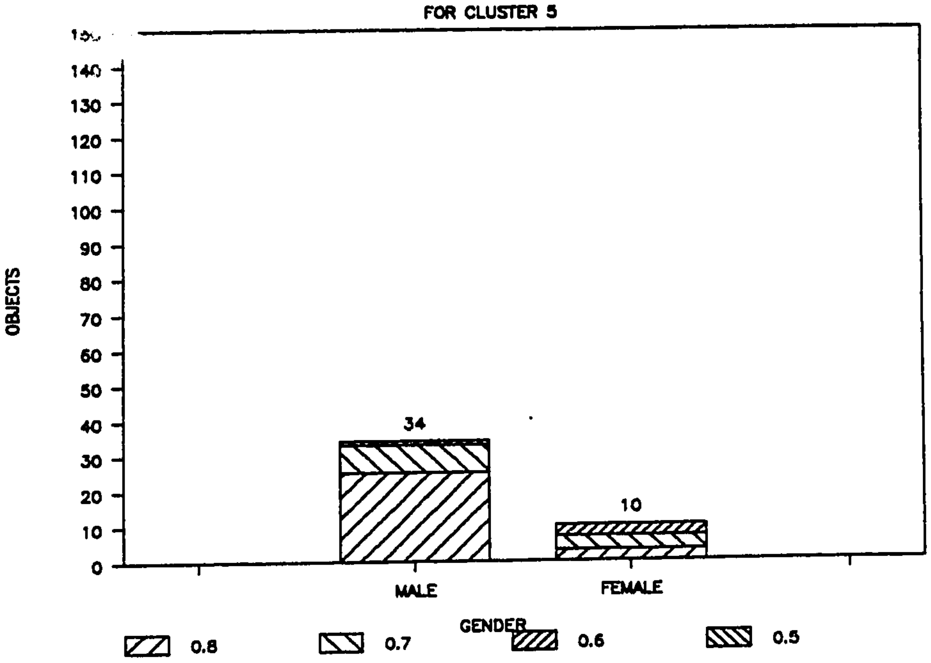

AGE DISTRIBUTION
FOR CLUSTER 4
OBJECTS
100
90
80
70
60
50
40
30
20
10
0
0
10
10
10
18
14
5
7
0
0
0
15–19
20–24
25–29
30–34
35–39
40–44
45–49
50–54
55–59
60–64
65–69
AGE CATEGORIES
0.8
0.7
0.6
0.5


JOB CATEGORY DISTRIBUTION
FOR CLUSTER 4
OBJECTS
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
0
0
0
3
71
1
2
3
4
5
JOB CATEGORIES
0.8
0.7
0.6
0.5

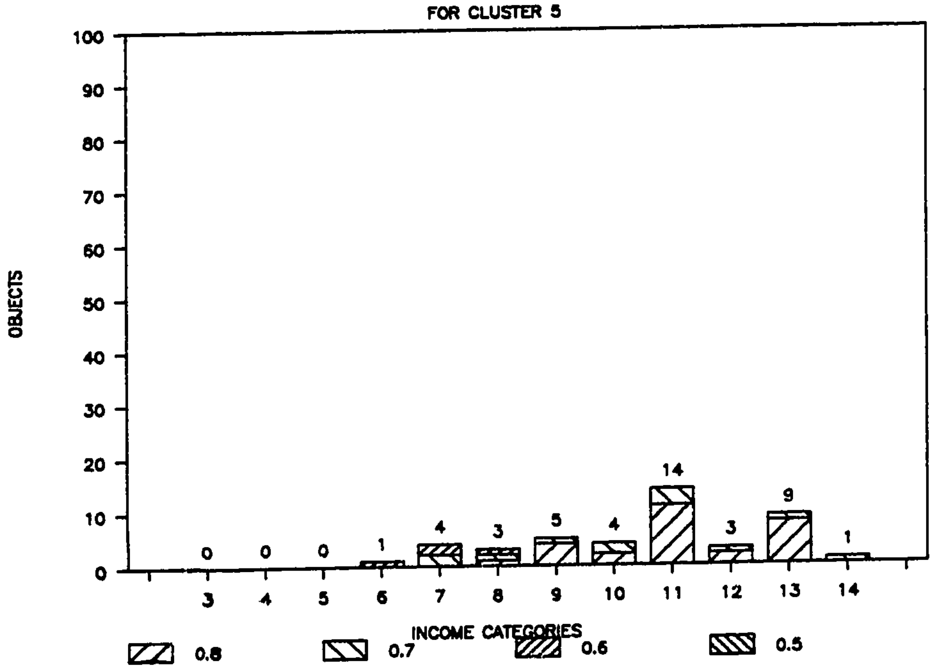

MODE DISTRIBUTION
FOR CLUSTER 4
OBJECTS
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
0
74
CAR
PUBLIC TRANSPORT
MODE
0.8
0.7
0.6
0.5

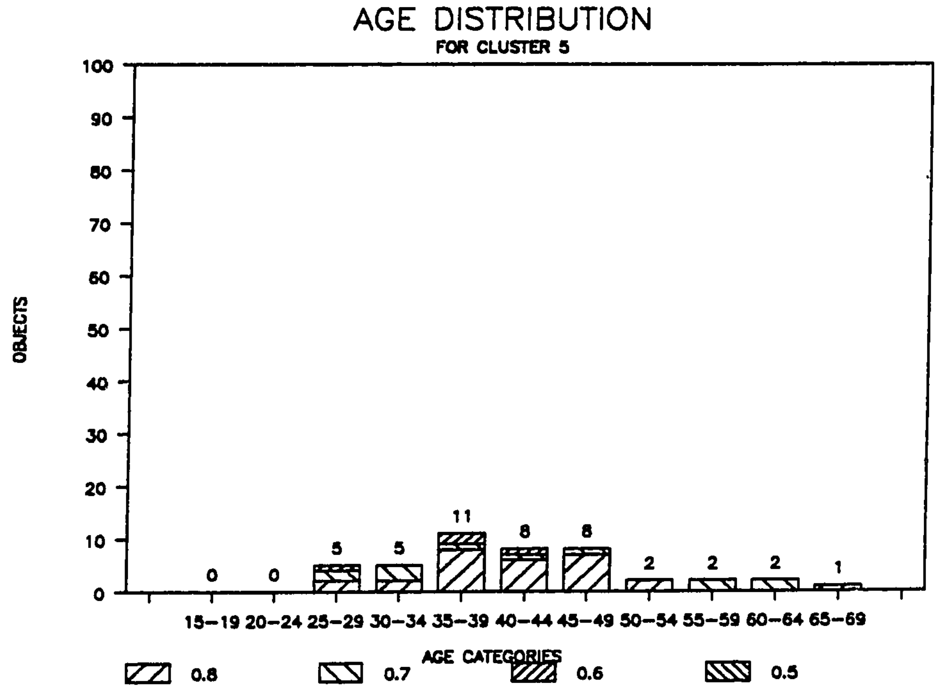

MODE FREQUENCY DISTRIBUTION
FOR CLUSTER 4
OBJECTS
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
0
0
9
65
1
2
3
4
MODE FREQUENCY CATEGORIES
0.8
0.7
0.6
0.5

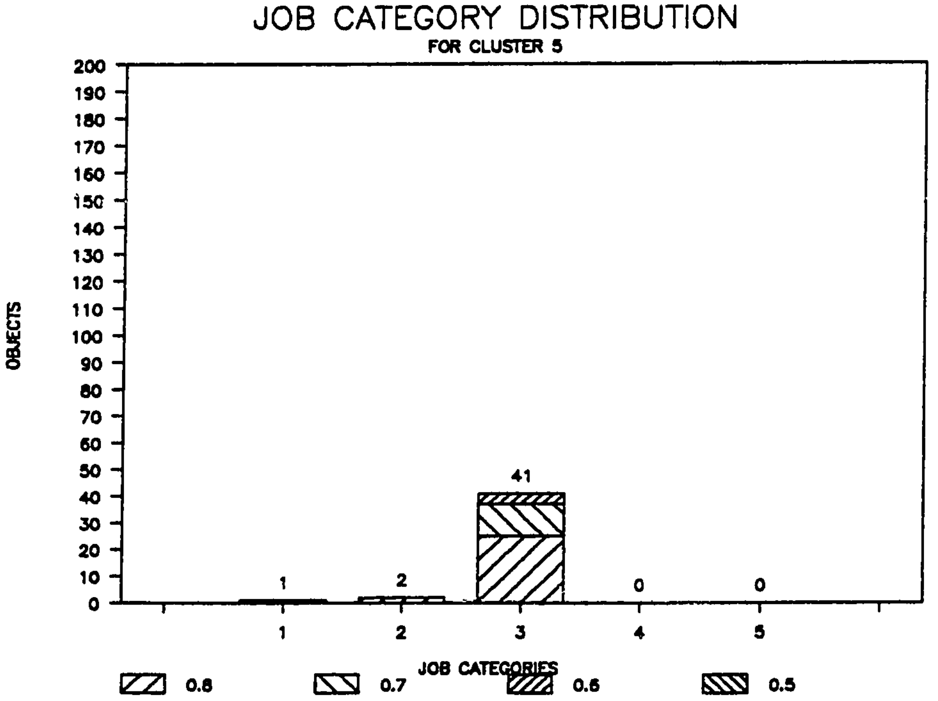
GENDER DISTRIBUTION
FOR CLUSTER 5

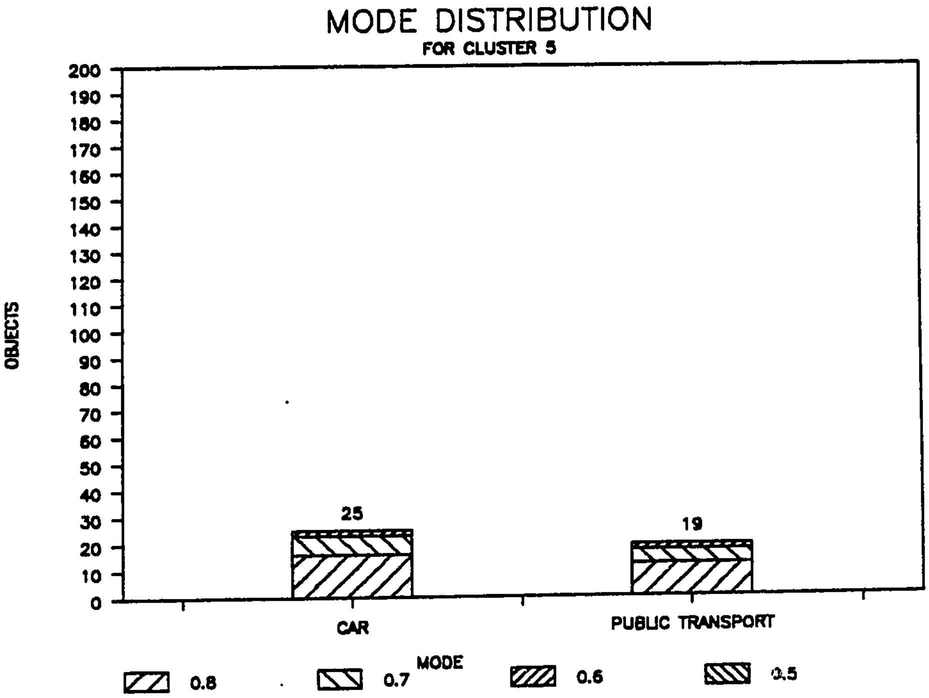


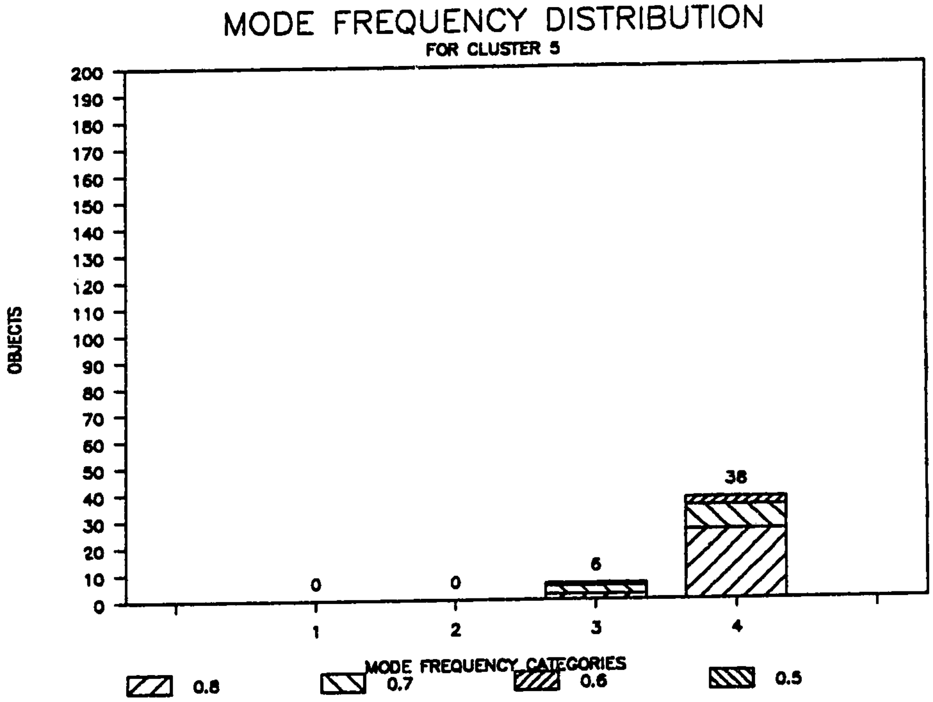
INCOME DISTRIBUTION
FOR CLUSTER 5

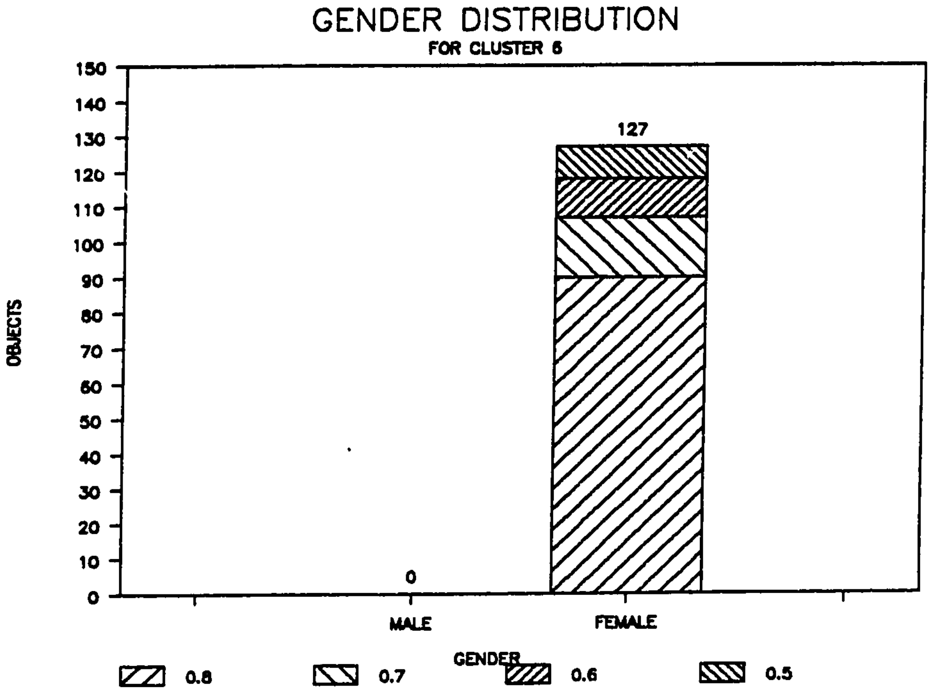


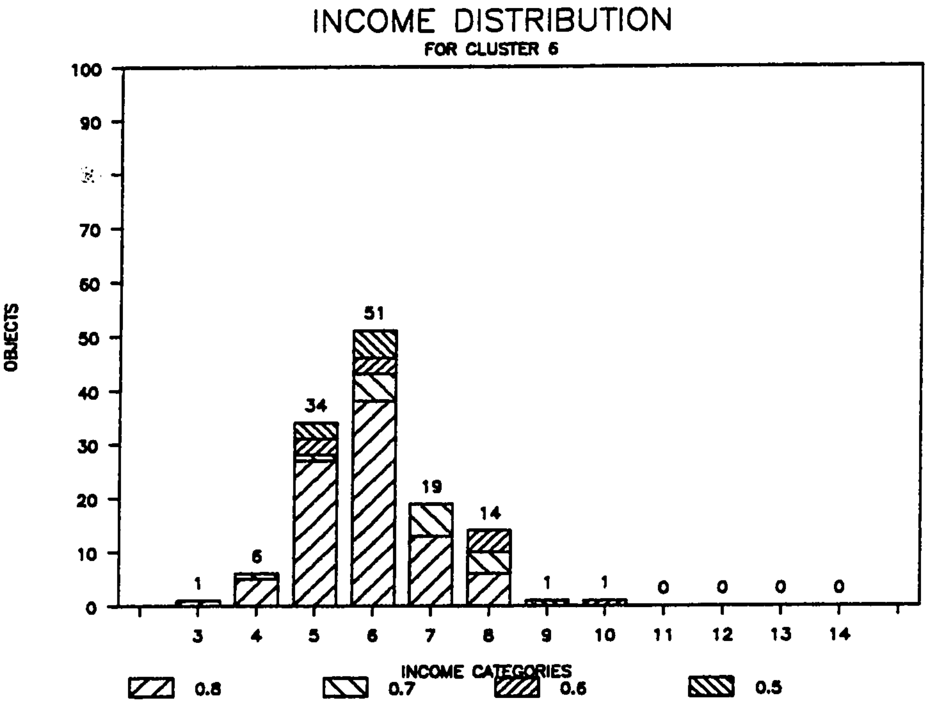

AGE DISTRIBUTION
FOR CLUSTER 5
OBJECTS
100
90
80
70
60
50
40
30
20
10
0
0
0
5
5
11
8
8
2
2
2
1
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
AGE CATEGORIES
0.8
0.7
0.6
0.5

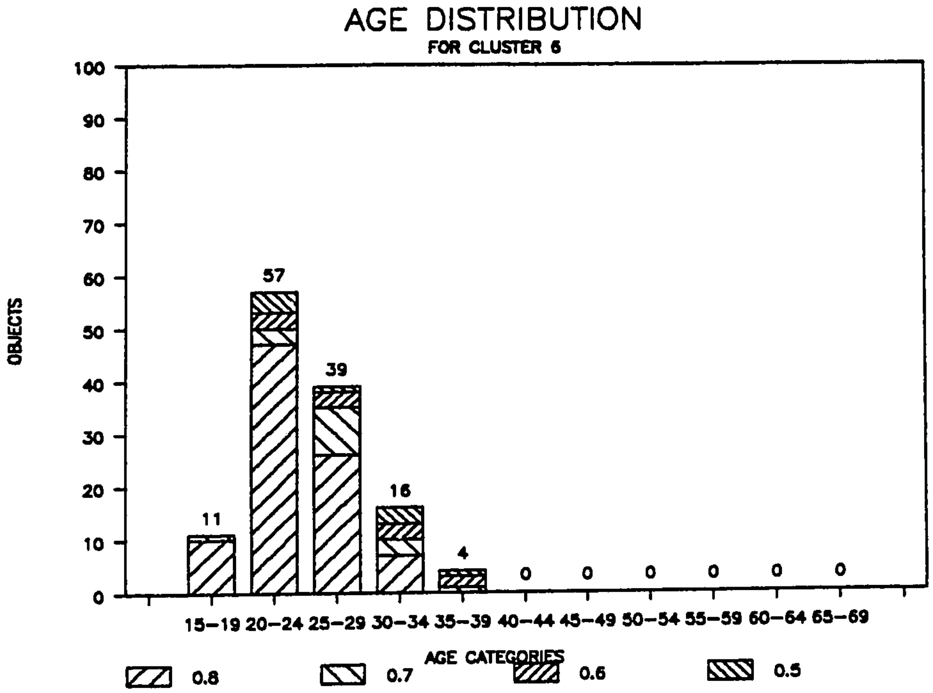

JOB CATEGORY DISTRIBUTION
FOR CLUSTER 5
OBJECTS
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
1
2
41
0
0
1
2
3
4
5
JOB CATEGORIES
0.8
0.7
0.6
0.5

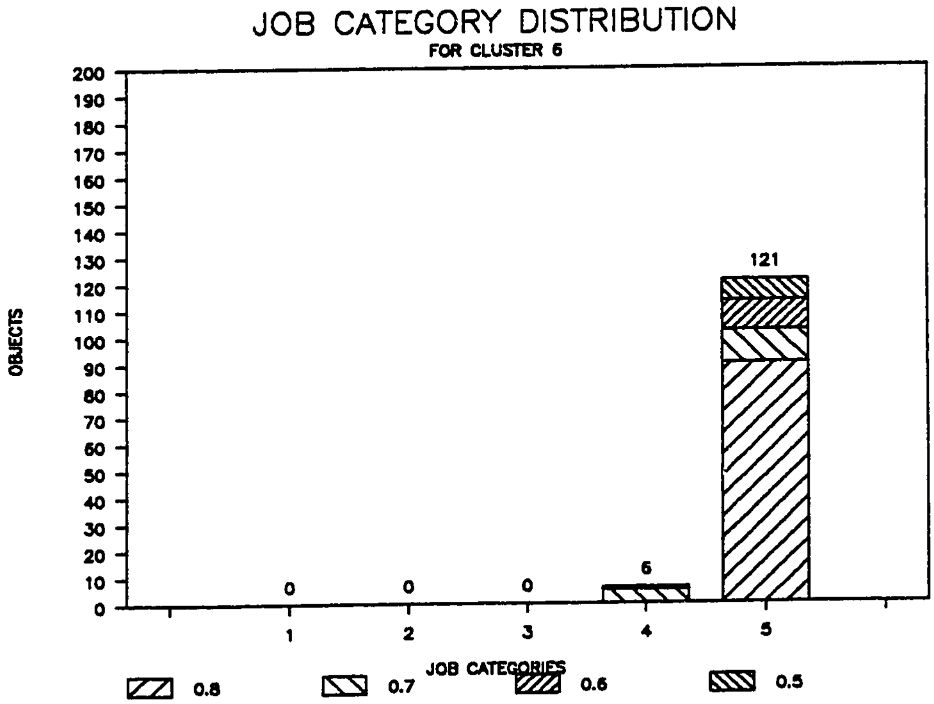

MODE DISTRIBUTION
FOR CLUSTER 5
OBJECTS
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
25
19
CAR
PUBLIC TRANSPORT
MODE
0.8
0.7
0.6
0.5

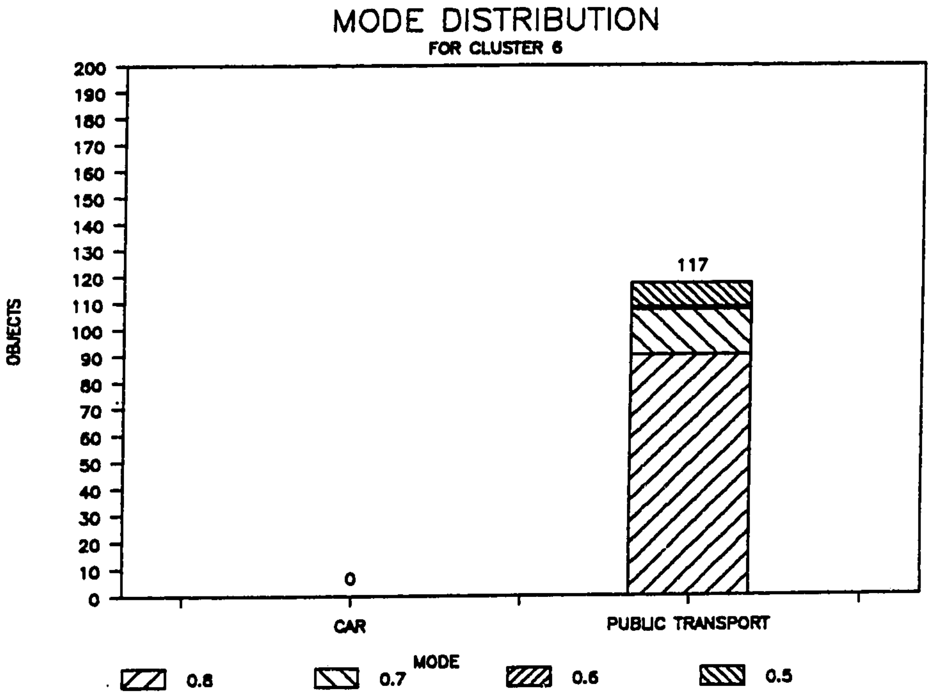

MODE FREQUENCY DISTRIBUTION
FOR CLUSTER 5
OBJECTS
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
0
0
6
38
1
2
3
4
MODE FREQUENCY CATEGORIES
0.8
0.7
0.6
0.5

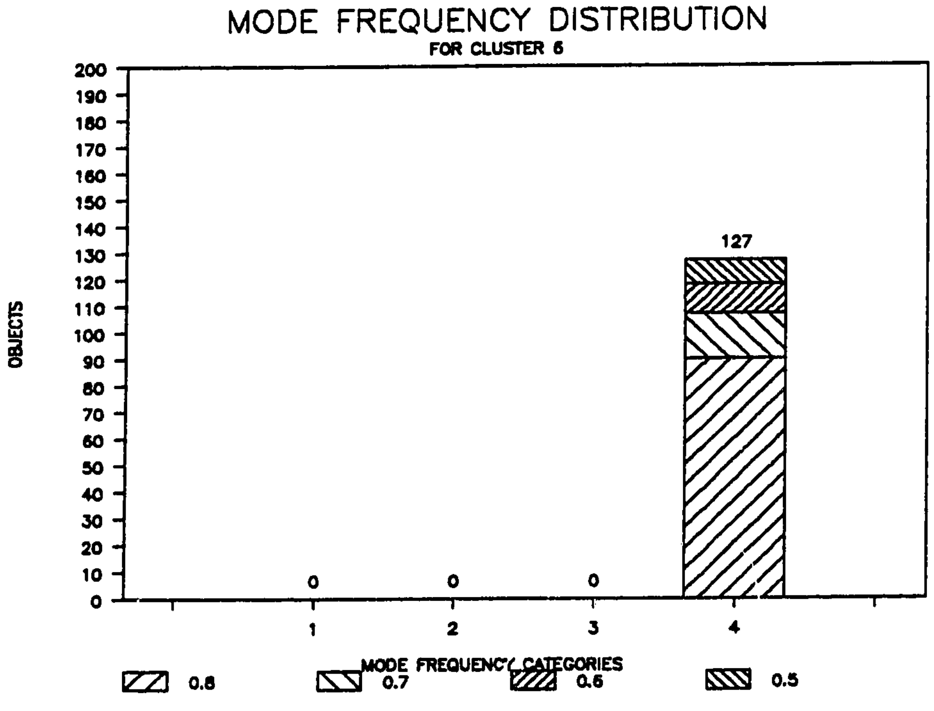

GENDER DISTRIBUTION
FOR CLUSTER 6
OBJECTS
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
0
127
MALE
FEMALE
GENDER
0.8
0.7
0.6
0.5


INCOME DISTRIBUTION
FOR CLUSTER 6
OBJECTS
100
90
70
60
50
40
30
20
10
0
1
6
34
51
19
14
1
1
0
0
0
0
3
4
5
6
7
8
9
10
11
12
13
14
INCOME CATEGORIES
0.8
0.7
0.6
0.5


AGE DISTRIBUTION
FOR CLUSTER 6
OBJECTS
0
10
20
30
40
50
60
70
80
90
100
11
57
39
16
4
0
0
0
0
0
0
15–19
20–24
25–29
30–34
35–39
40–44
45–49
50–54
55–59
60–64
65–69
AGE CATEGORIES
0.8
0.7
0.6
0.5


JOB CATEGORY DISTRIBUTION
FOR CLUSTER 6
OBJECTS
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
0
0
0
6
121
1
2
3
4
5
JOB CATEGORIES
0.8
0.7
0.6
0.5

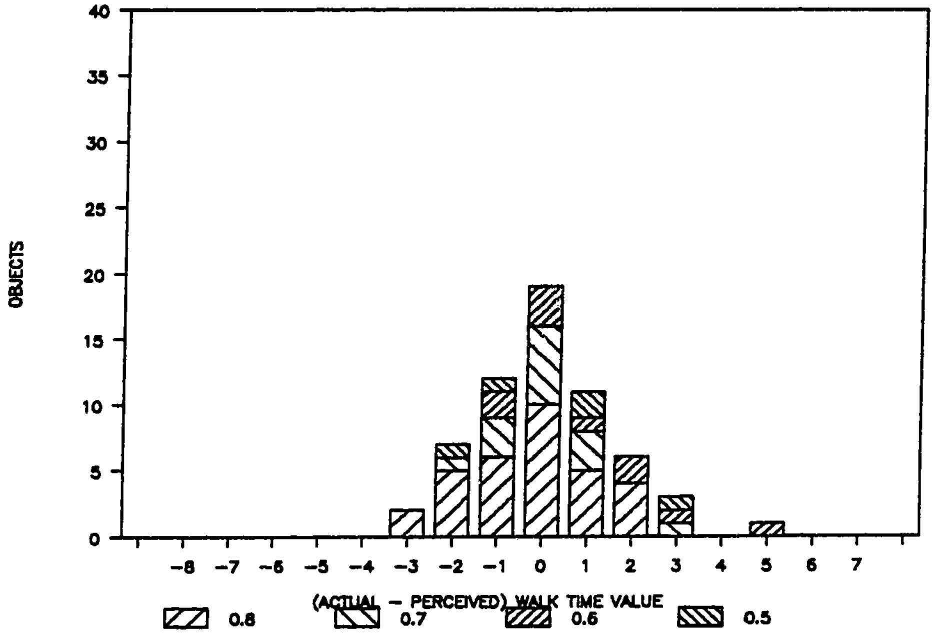

MODE DISTRIBUTION
FOR CLUSTER 6
OBJECTS
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
117
0
CAR
PUBLIC TRANSPORT
MODE
0.8
0.7
0.6
0.5


MODE FREQUENCY DISTRIBUTION
FOR CLUSTER 6
OBJECTS
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
127
0
0
0
1
2
3
4
MODE FREQUENCY CATEGORIES
0.8
0.7
0.6
0.5

Appendix D
Walk and Wait Time Ratio and Walk and Wait Time Difference Distributions for the Six Cluster Solution Cluster Cores

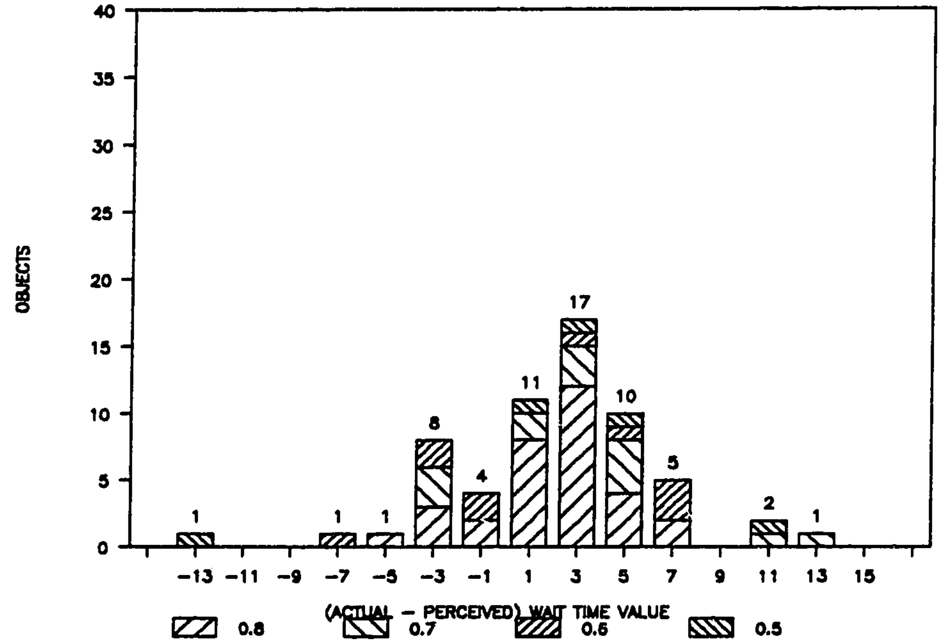
ACTUAL-PERCEIVED WALK TIME DISTRIBUTION

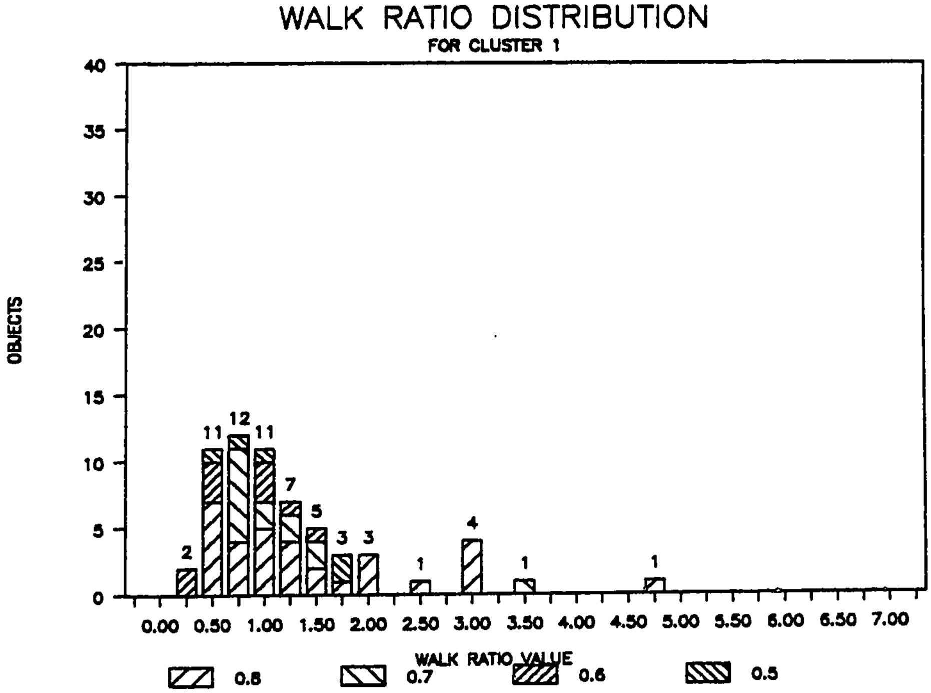
FOR CLUSTER 1

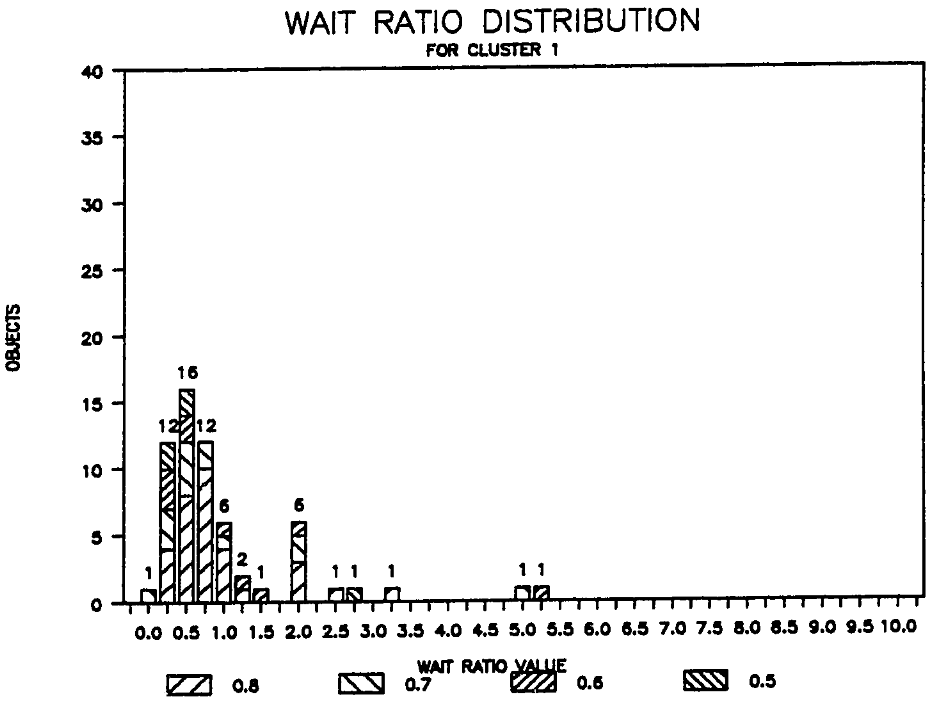


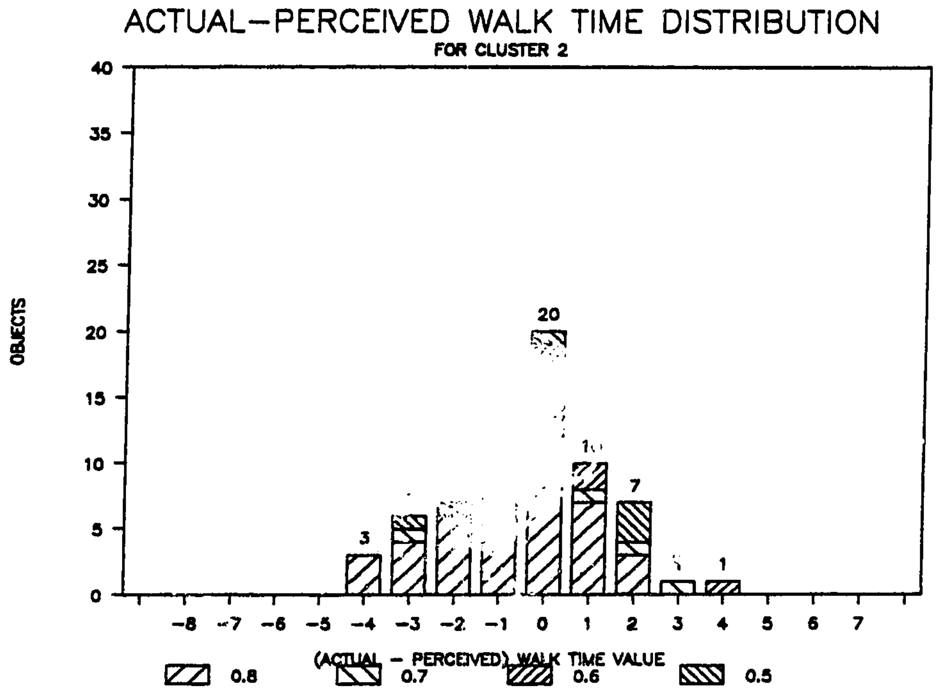
ACTUAL-PERCEIVED WAIT TIME DISTRIBUTION

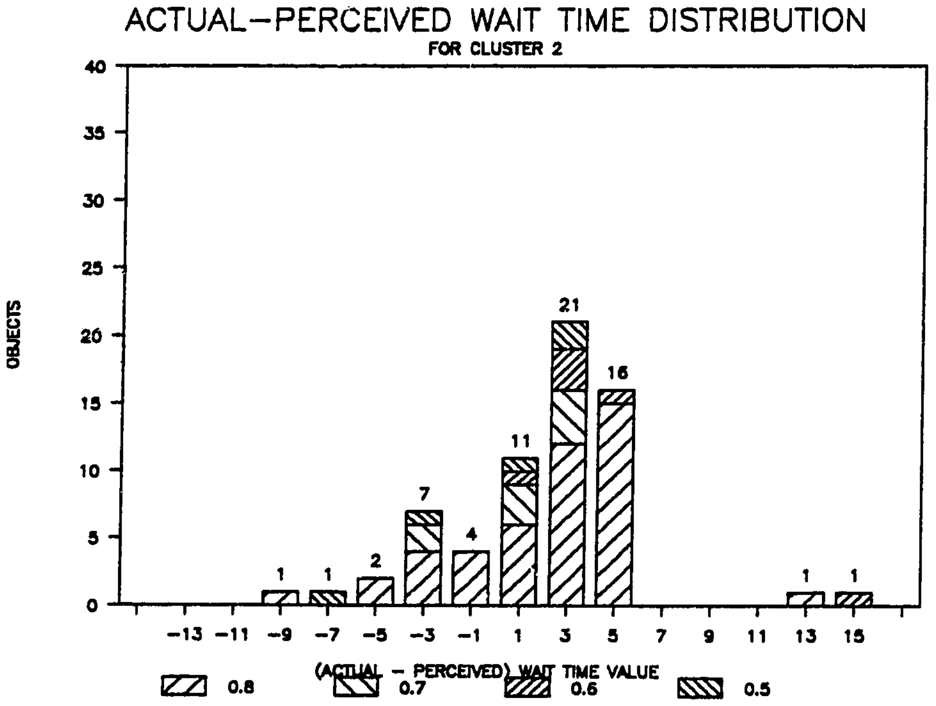
FOR CLUSTER 1

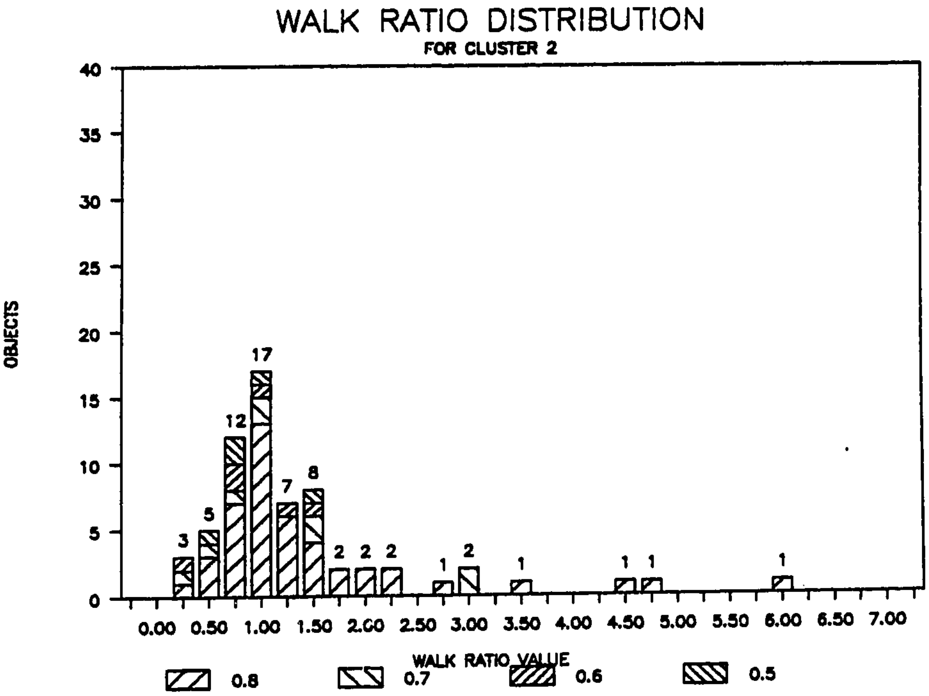


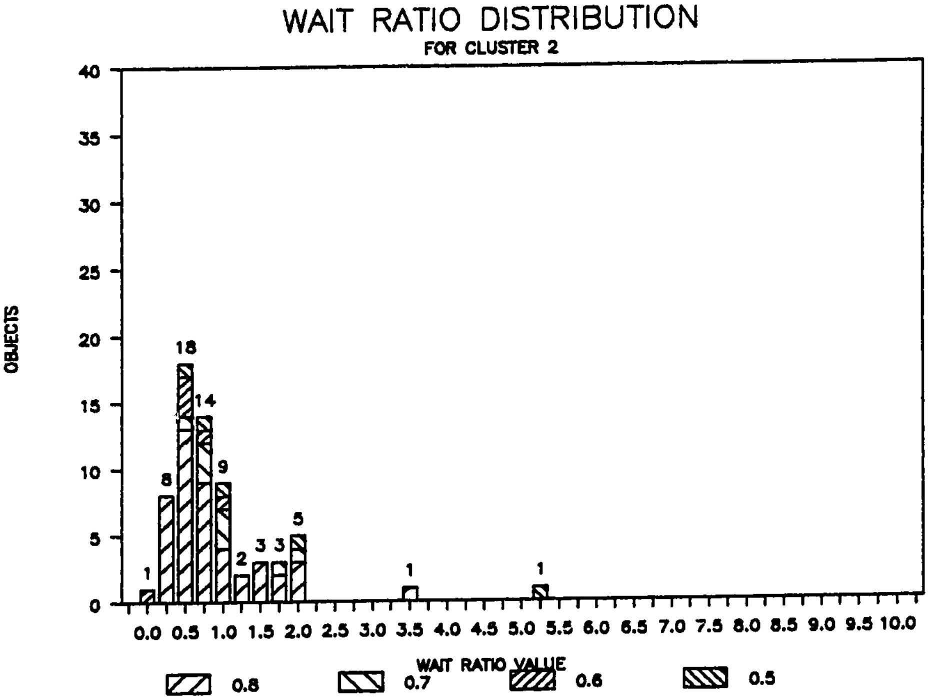

WALK RATIO DISTRIBUTION
FOR CLUSTER 1
OBJECTS
0
5
10
15
20
25
30
35
40
2
11
12
11
7
5
3
3
1
4
1
1
0.00
0.50
1.00
1.50
2.00
2.50
3.00
3.50
4.00
4.50
5.00
5.50
6.00
6.50
7.00
WALK RATIO VALUE
0.8
0.7
0.6
0.5


WAIT RATIO DISTRIBUTION
FOR CLUSTER 1
OBJECTS
0
5
10
15
20
25
30
35
40
1
12
16
12
6
2
1
6
1
1
1
1
1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
9.5
10.0
WAIT RATIO VALUE
0.8
0.7
0.6
0.5

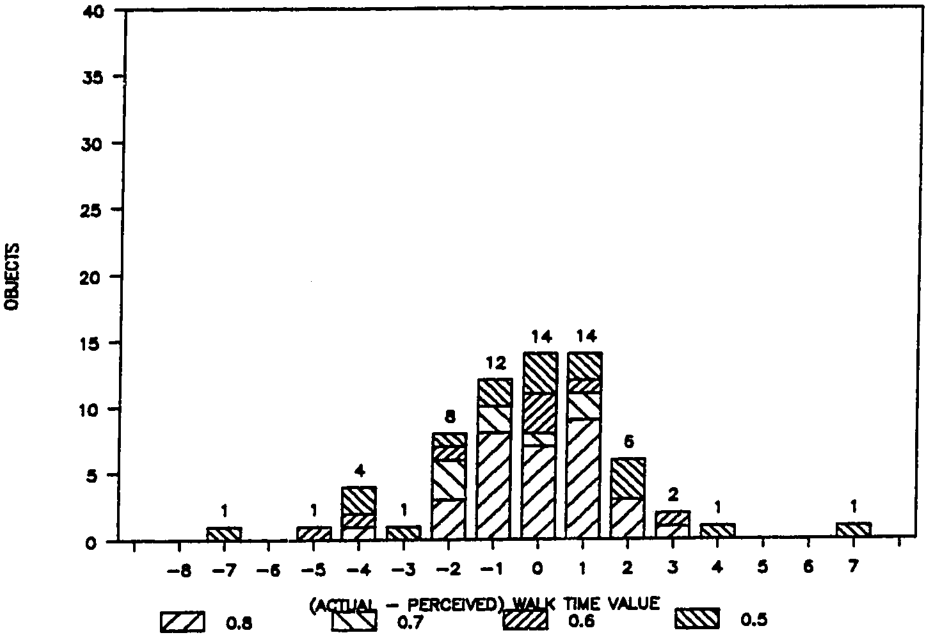

ACTUAL-PERCEIVED WALK TIME DISTRIBUTION
FOR CLUSTER 2
OBJECTS
40
35
30
25
20
15
10
5
0
20
10
7
3
1
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
(ACTUAL - PERCEIVED) WALK TIME VALUE
0.8
0.7
0.6
0.5


ACTUAL-PERCEIVED WAIT TIME DISTRIBUTION
FOR CLUSTER 2
OBJECTS
40
35
30
25
20
15
10
5
0
1
1
2
7
4
11
21
16
1
1
-13
-11
-9
-7
-5
-3
-1
1
3
5
7
9
11
13
15
(ACTUAL - PERCEIVED) WAIT TIME VALUE
0.8
0.7
0.6
0.5

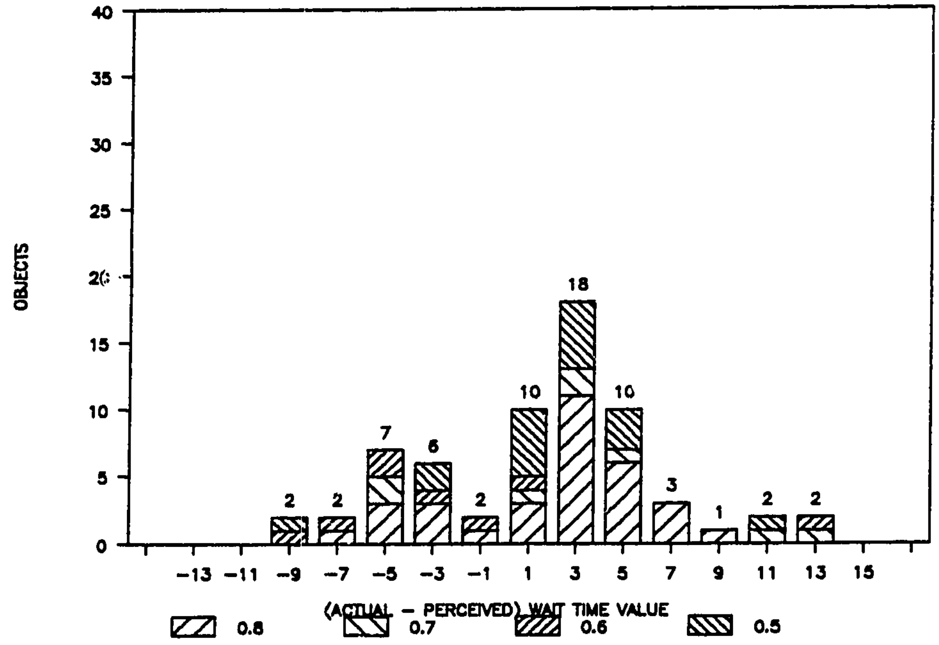

WALK RATIO DISTRIBUTION
FOR CLUSTER 2
OBJECTS
40
35
30
25
20
15
10
5
0
3
5
12
17
7
8
2
2
2
1
2
1
1
1
1
0.00
0.50
1.00
1.50
2.00
2.50
3.00
3.50
4.00
4.50
5.00
5.50
6.00
6.50
7.00
WALK RATIO VALUE
0.8
0.7
0.6
0.5

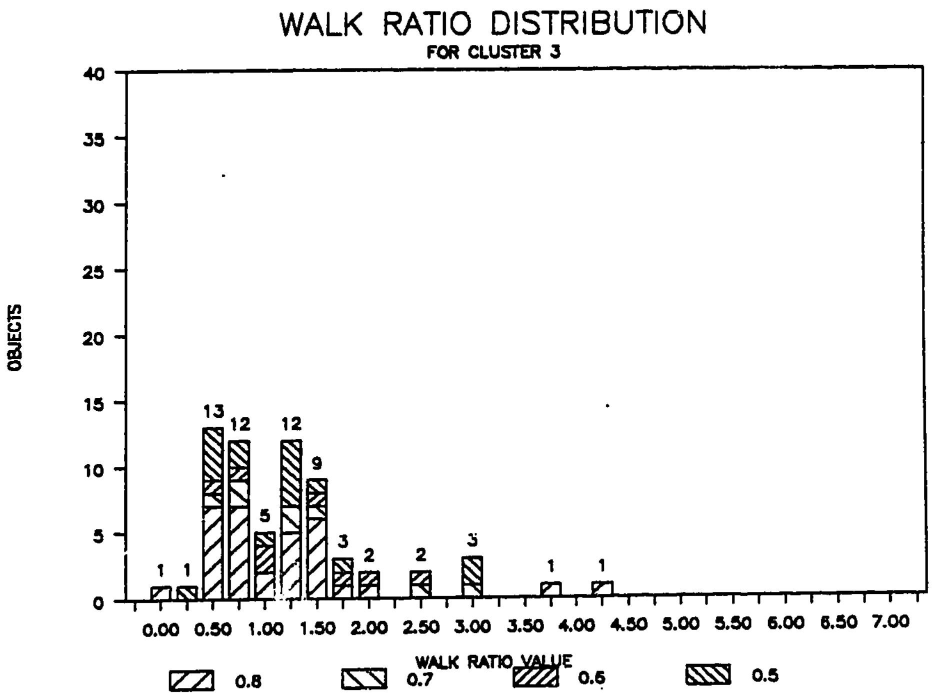

WAIT RATIO DISTRIBUTION
FOR CLUSTER 2
OBJECTS
40
35
30
25
20
15
10
5
0
1
8
18
14
9
2
3
3
5
1
1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
9.5
10.0
WAIT RATIO VALUE
0.8
0.7
0.6
0.5

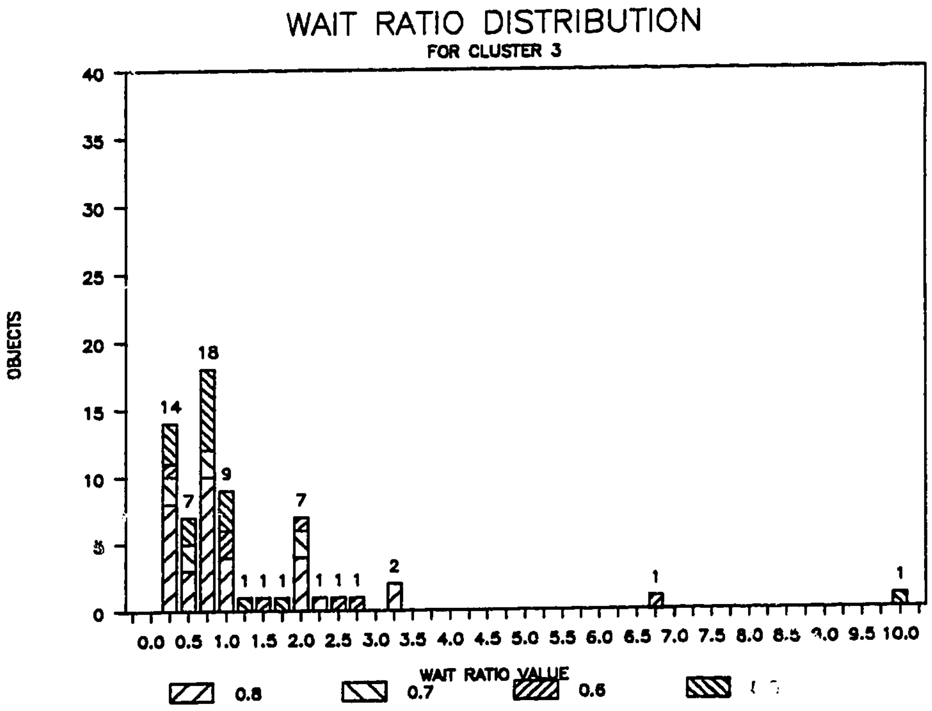
ACTUAL–PERCEIVED WALK TIME DISTRIBUTION
FOR CLUSTER 3

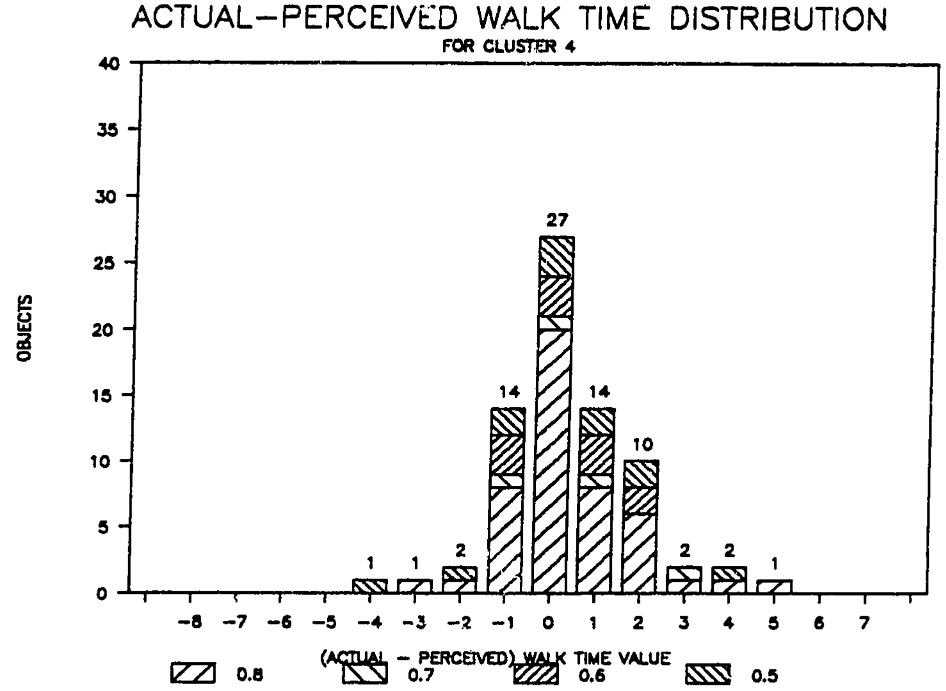


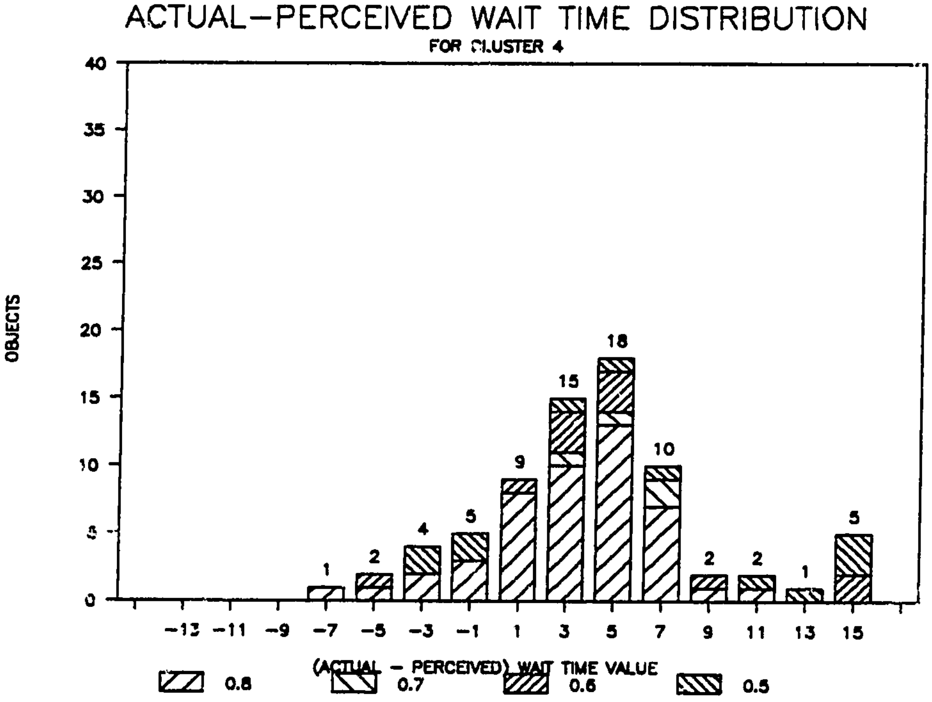
ACTUAL–PERCEIVED WAIT TIME DISTRIBUTION
FOR CLUSTER 3

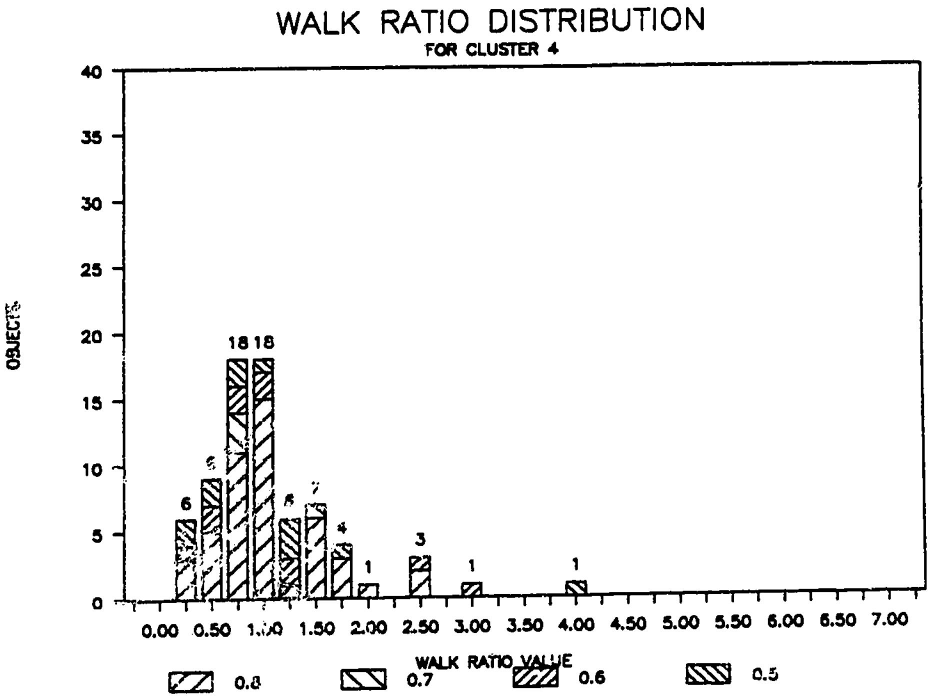


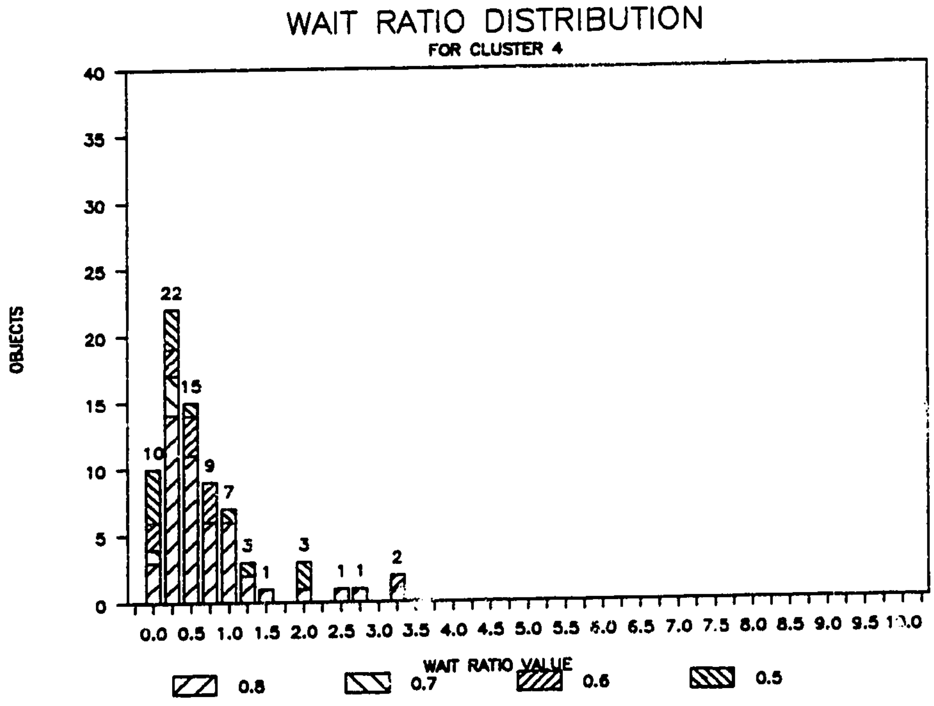

WALK RATIO DISTRIBUTION
FOR CLUSTER 3
OBJECTS
40
35
30
25
20
15
10
5
0
1
1
13
12
5
12
9
3
2
2
3
1
1
0.00
0.50
1.00
1.50
2.00
2.50
3.00
3.50
4.00
4.50
5.00
5.50
6.00
6.50
7.00
WALK RATIO VALUE
▨ 0.8
▧ 0.7
▨ 0.6
▧ 0.5


WAIT RATIO DISTRIBUTION
FOR CLUSTER 3
OBJECTS
40
35
30
25
20
15
10
5
0
14
7
18
9
1
1
1
7
1
1
1
2
1
1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
9.5
10.0
WAIT RATIO VALUE
▨ 0.8
▧ 0.7
▨ 0.6

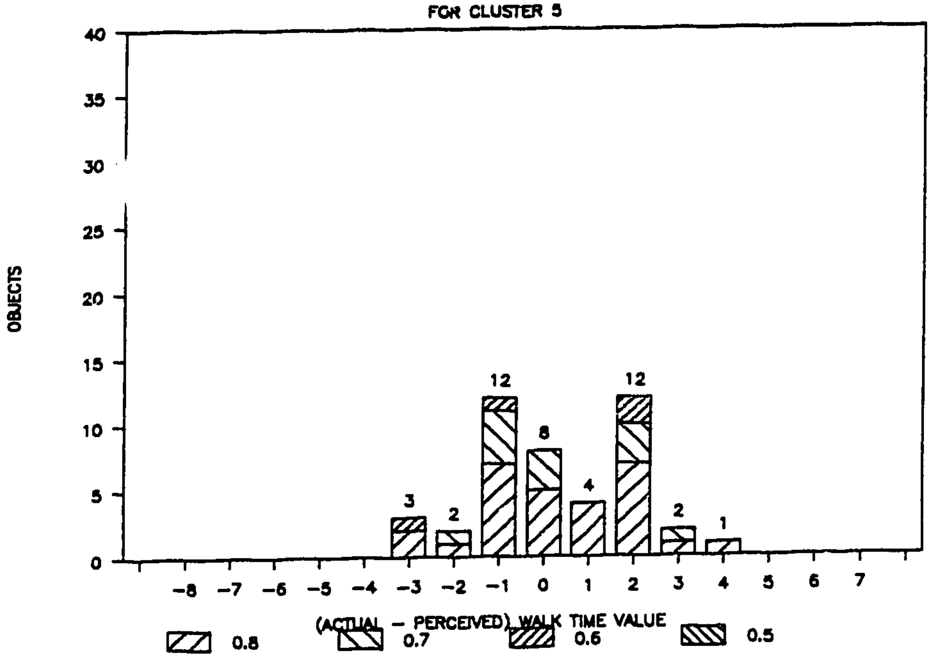

ACTUAL–PERCEIVED WALK TIME DISTRIBUTION
FOR CLUSTER 4
OBJECTS
40
35
30
25
20
15
10
5
0
27
14
14
10
1
1
2
2
2
1
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
(ACTUAL – PERCEIVED) WALK TIME VALUE
0.8
0.7
0.6
0.5


ACTUAL–PERCEIVED WAIT TIME DISTRIBUTION
FOR CLUSTER 4
OBJECTS
40
35
30
25
20
15
10
5
0
18
15
10
9
5
5
4
2
1
2
2
1
-13
-11
-9
-7
-5
-3
-1
1
3
5
7
9
11
13
15
(ACTUAL – PERCEIVED) WAIT TIME VALUE
0.8
0.7
0.6
0.5

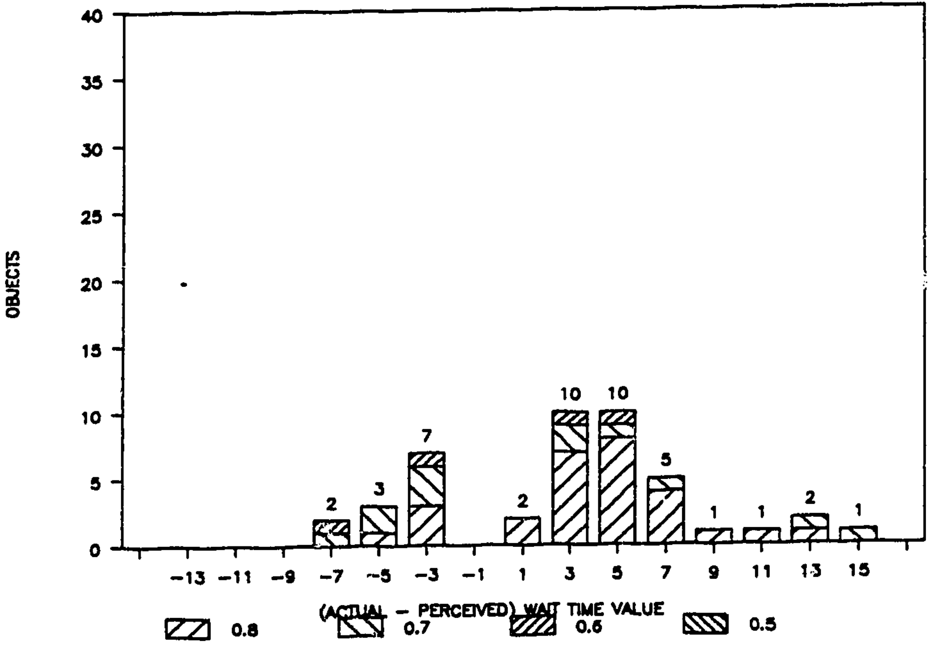

WALK RATIO DISTRIBUTION
FOR CLUSTER 4
OBJECTS
40
35
30
25
20
15
10
5
0
6
9
18
18
6
7
4
1
3
1
1
0.00
0.50
1.00
1.50
2.00
2.50
3.00
3.50
4.00
4.50
5.00
5.50
6.00
6.50
7.00
WALK RATIO VALUE
0.8
0.7
0.6
0.5

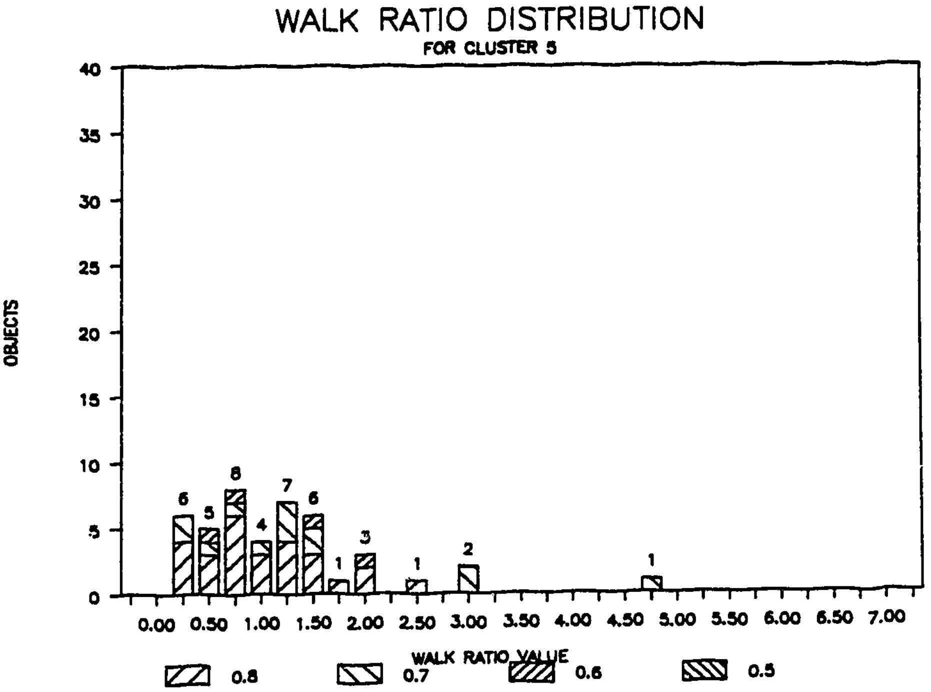

WAIT RATIO DISTRIBUTION
FOR CLUSTER 4
OBJECTS
40
35
30
25
20
15
10
5
0
10
22
15
9
7
3
1
3
1
1
2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
9.5
10.0
WAIT RATIO VALUE
0.8
0.7
0.6
0.5

ACTUAL–PERCEIVED WALK TIME DISTRIBUTION
FOR CLUSTER 5

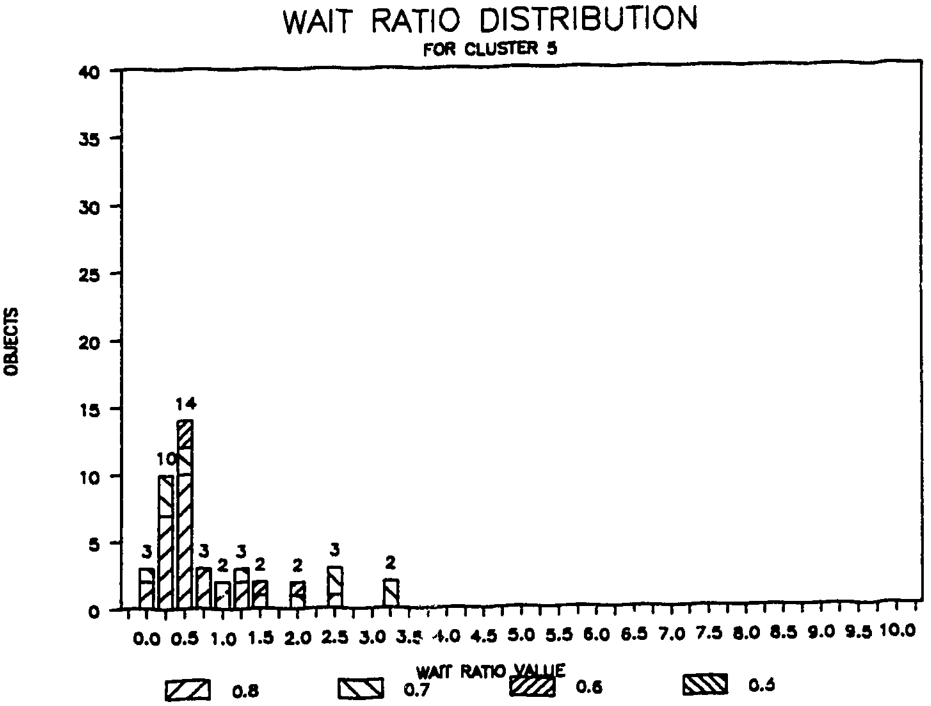


ACTUAL–PERCEIVED WAIT TIME DISTRIBUTION
FOR CLUSTER 5

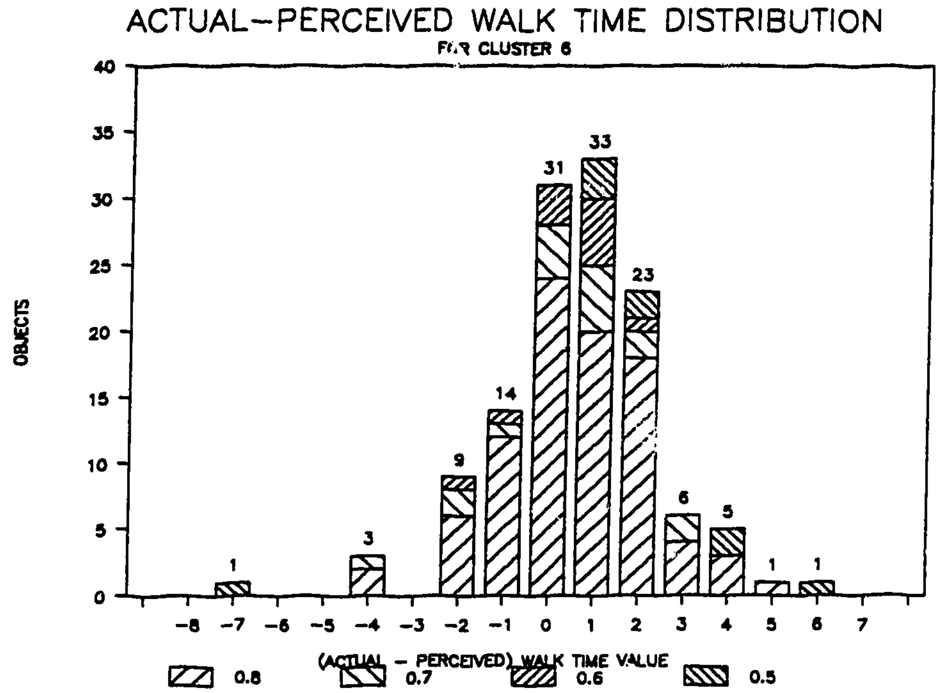


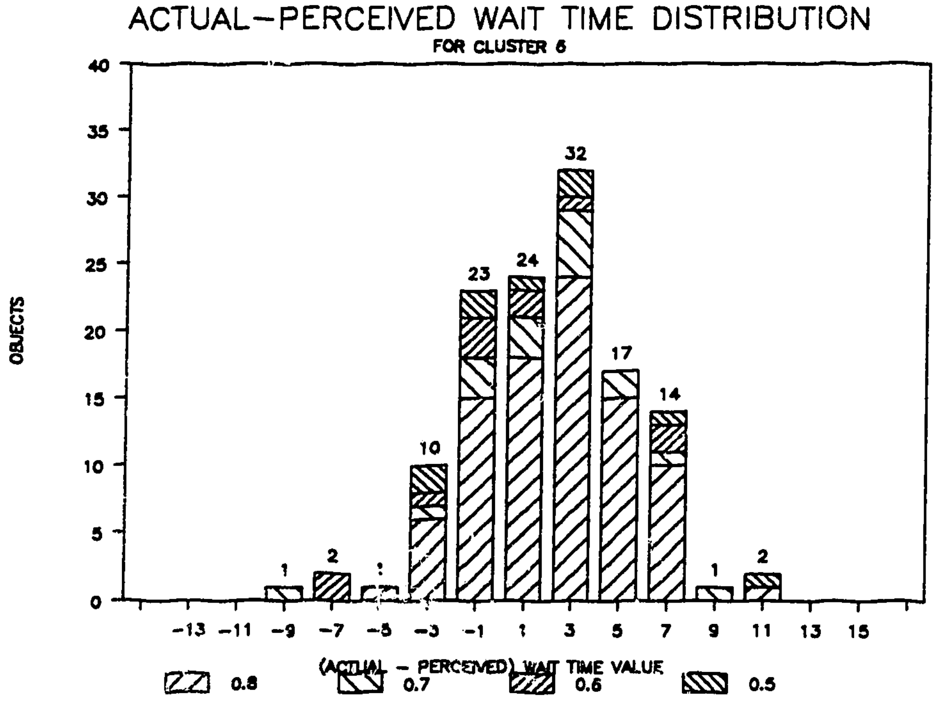

WALK RATIO DISTRIBUTION

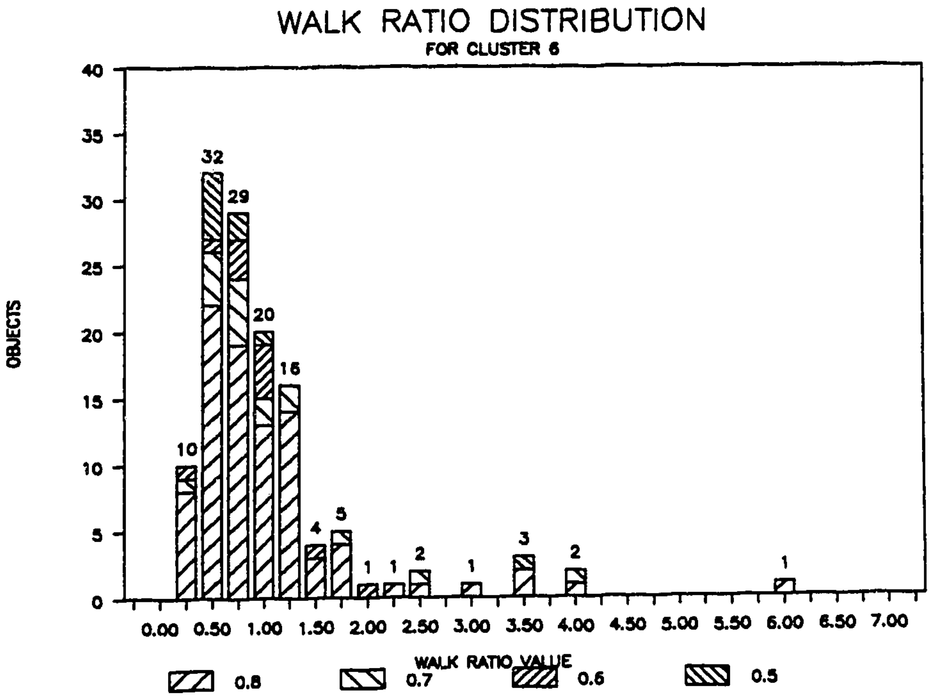
FOR CLUSTER 5

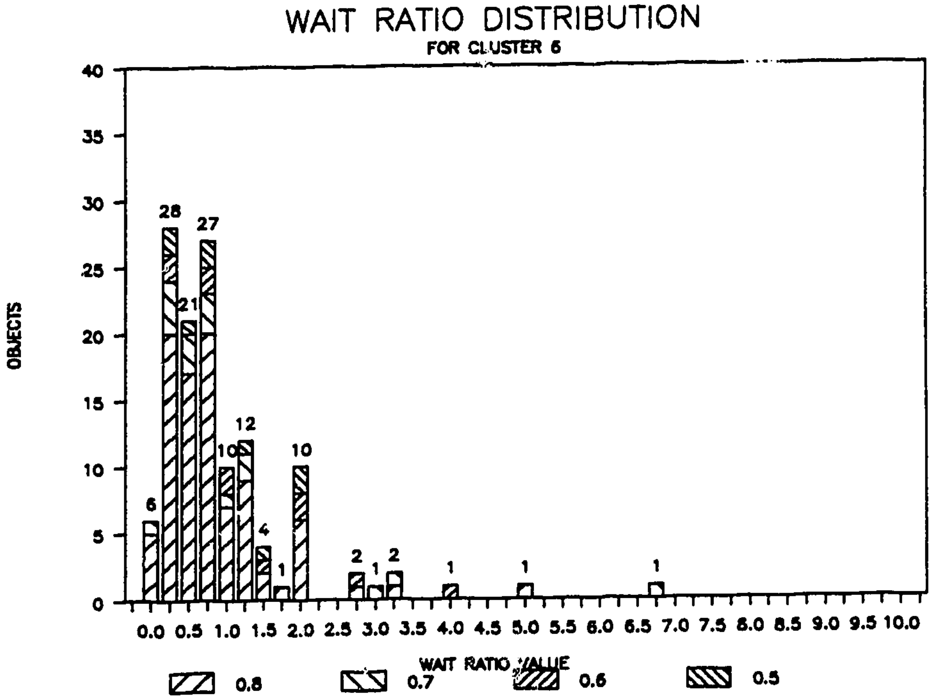

WAIT RATIO DISTRIBUTION

FOR CLUSTER 5

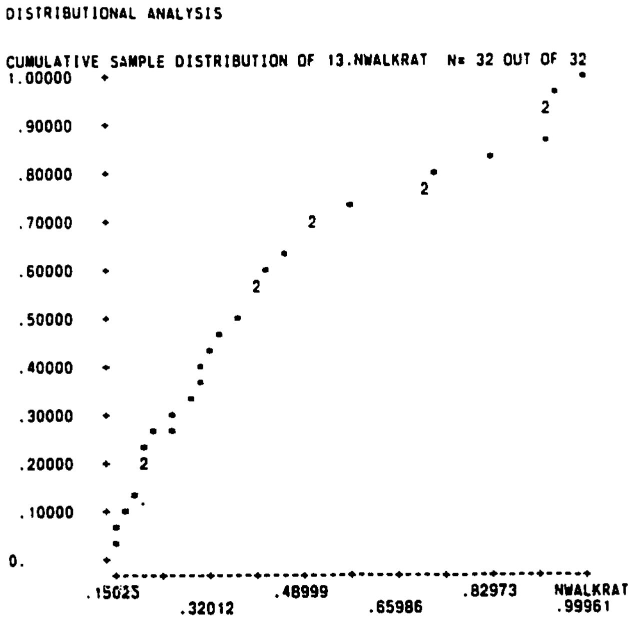

ACTUAL–PERCEIVED WALK TIME DISTRIBUTION
FOR CLUSTER 6
OBJECTS
40
35
30
25
20
15
10
5
0
1
3
9
14
31
33
23
6
5
1
1
–8
–7
–6
–5
–4
–3
–2
–1
0
1
2
3
4
5
6
7
(ACTUAL – PERCEIVED) WALK TIME VALUE
0.8
0.7
0.6
0.5


ACTUAL–PERCEIVED WAIT TIME DISTRIBUTION
FOR CLUSTER 6
OBJECTS
40
35
30
25
20
15
10
5
0
1
2
1
10
23
24
32
17
14
1
2
–13
–11
–9
–7
–5
–3
–1
1
3
5
7
9
11
13
15
(ACTUAL – PERCEIVED) WAIT TIME VALUE
0.8
0.7
0.6
0.5

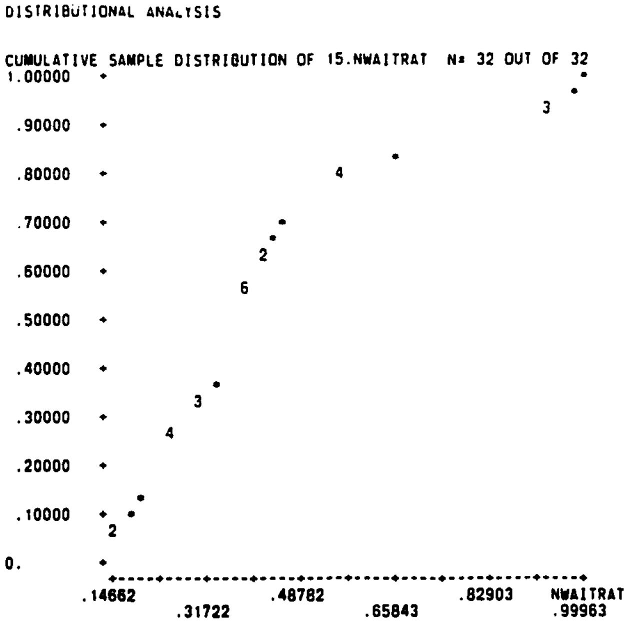

WALK RATIO DISTRIBUTION
FOR CLUSTER 6
OBJECTS
40
35
30
25
20
15
10
5
0
10
32
29
20
16
4
5
1
1
2
1
3
2
1
0.00 0.50 1.00 1.50 2.00 2.50 3.00 3.50 4.00 4.50 5.00 5.50 6.00 6.50 7.00
WALK RATIO VALUE
0.8
0.7
0.6
0.5


WAIT RATIO DISTRIBUTION
FOR CLUSTER 6
OBJECTS
40
35
30
25
20
15
10
5
0
6
28
21
27
10
12
4
1
10
2
1
2
1
1
1
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0 8.5 9.0 9.5 10.0
WAIT RATIO VALUE
0.8
0.7
0.6
0.5

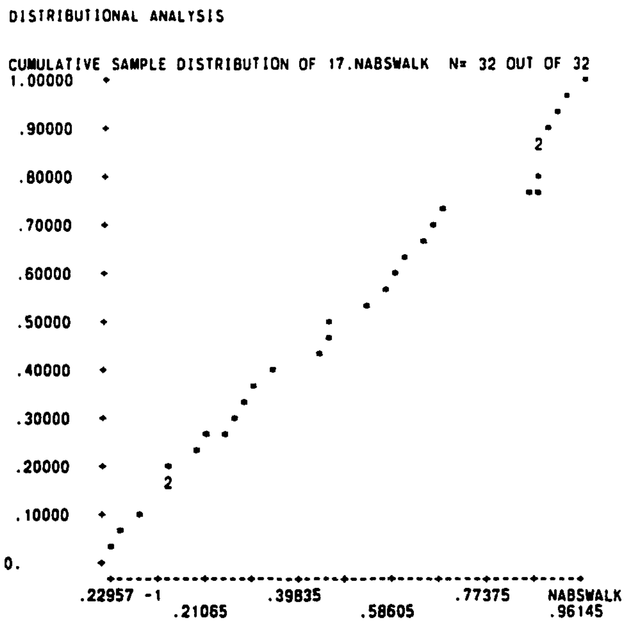
Appendix E
Normal Probability Plots for the Walk and Wait Time Ratio and Walk and Wait Time Difference Distributions for the Six Cluster Solution Cluster Cores



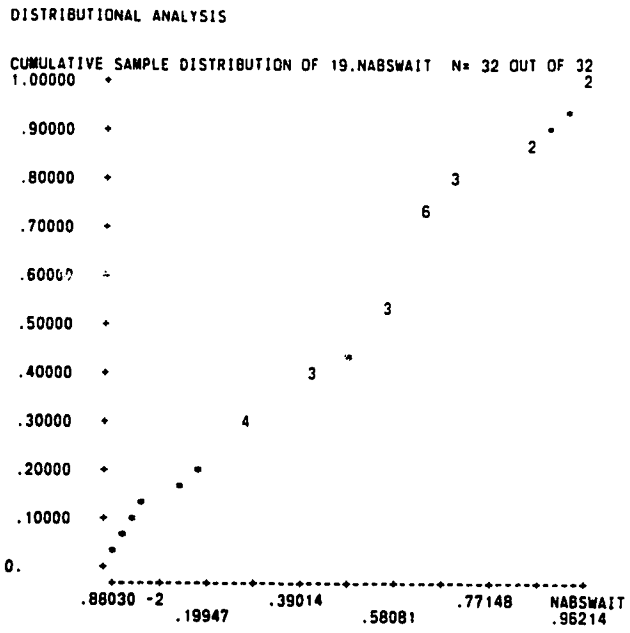
Cluster 1 Walk Time Ratio Normal Probability Plot



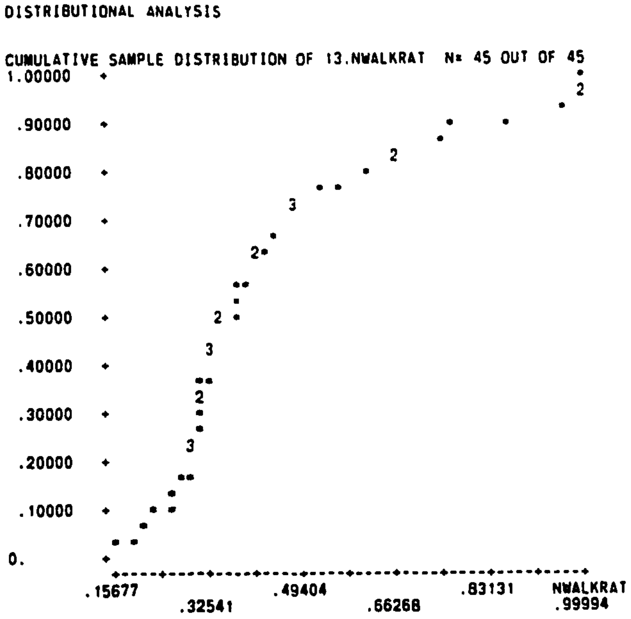
Cluster 1 Wait Time Ratio Normal Probability Plot



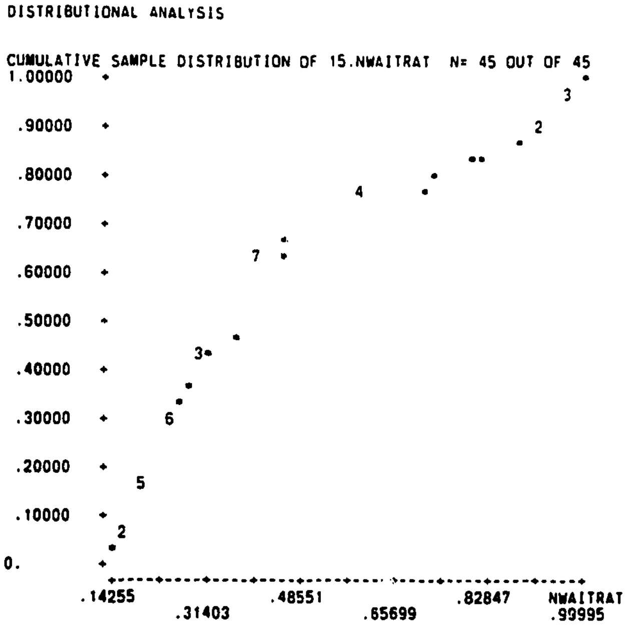
Cluster 1 Walk Time Difference Normal Probability Plot



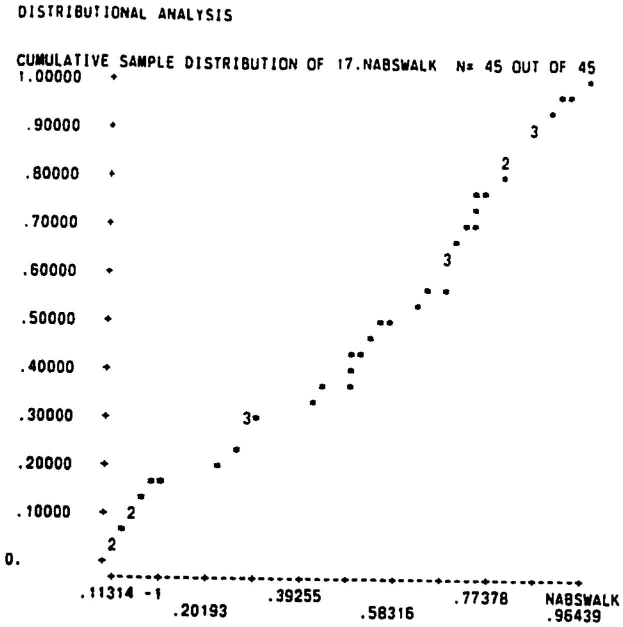
Cluster 1 Wait Time Difference Normal Probability Plot



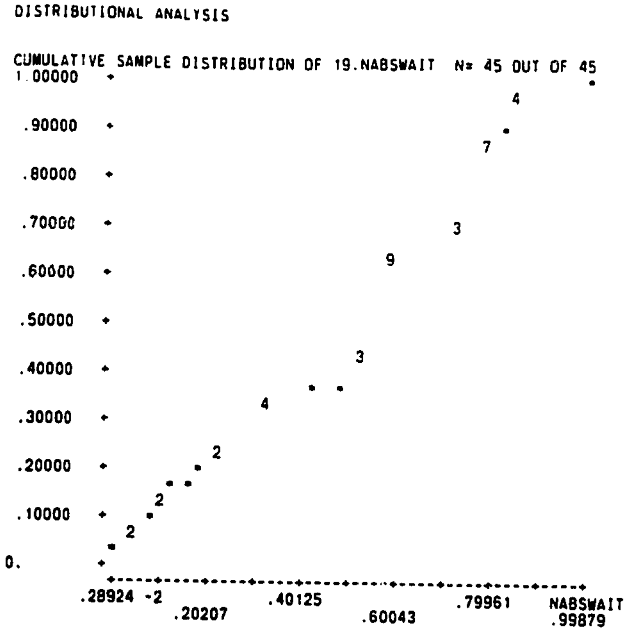
Cluster 2 Walk Time Ratio Normal Probability Plot



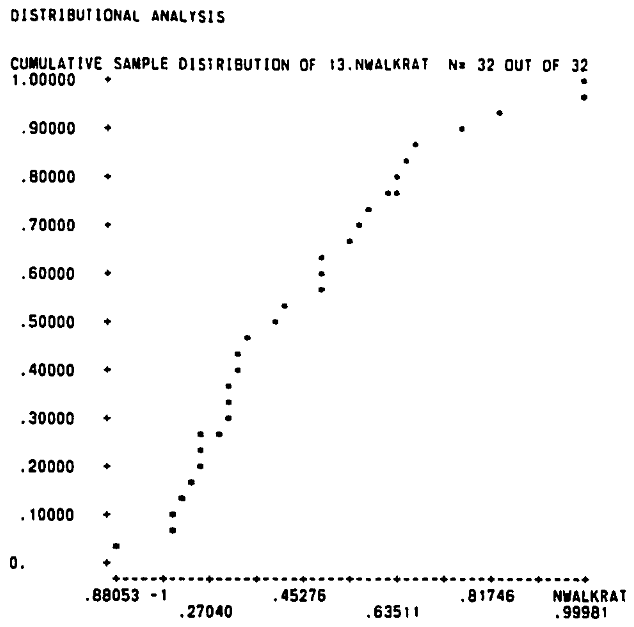
Cluster 2 Wait Time Ratio Normal Probability Plot



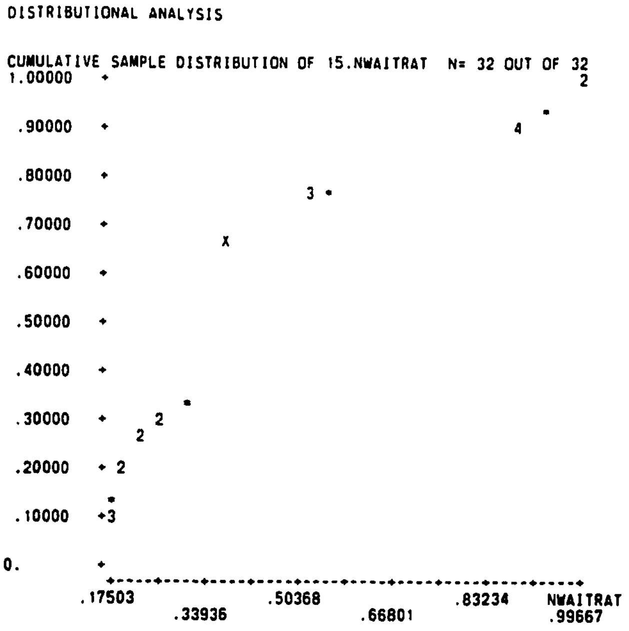
Cluster 2 Walk Time Difference Normal Probability Plot



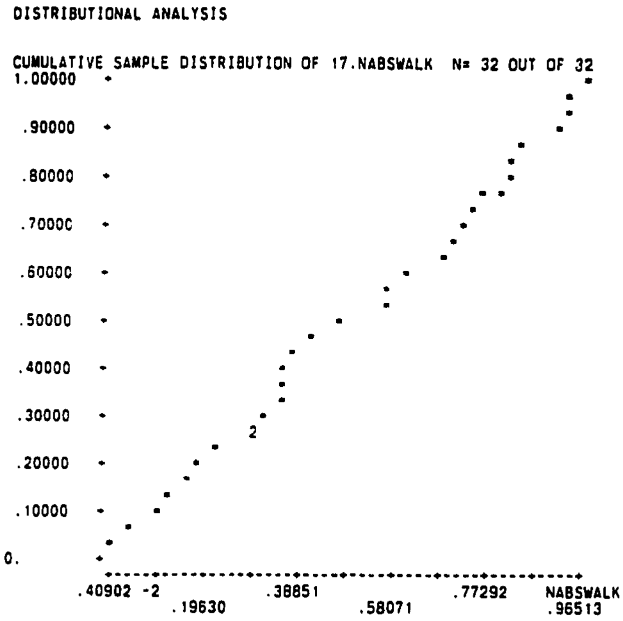
Cluster 2 Wait Time Difference Normal Probability Plot



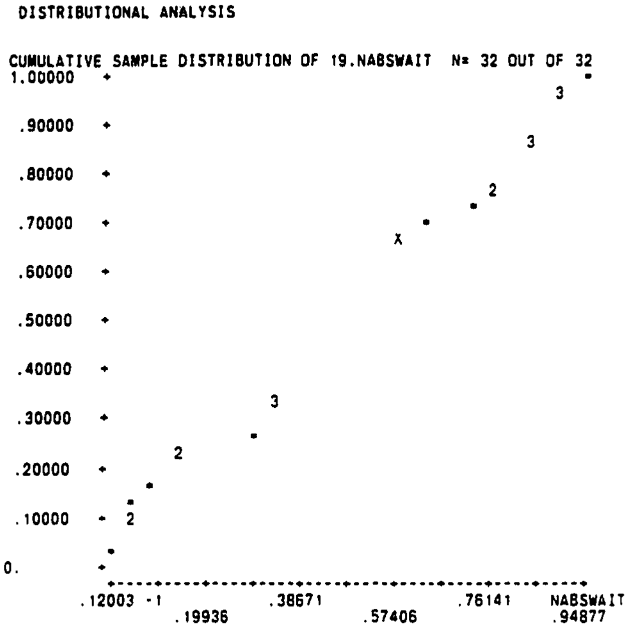
Cluster 3 Walk Time Ratio Normal Probabilility Plot



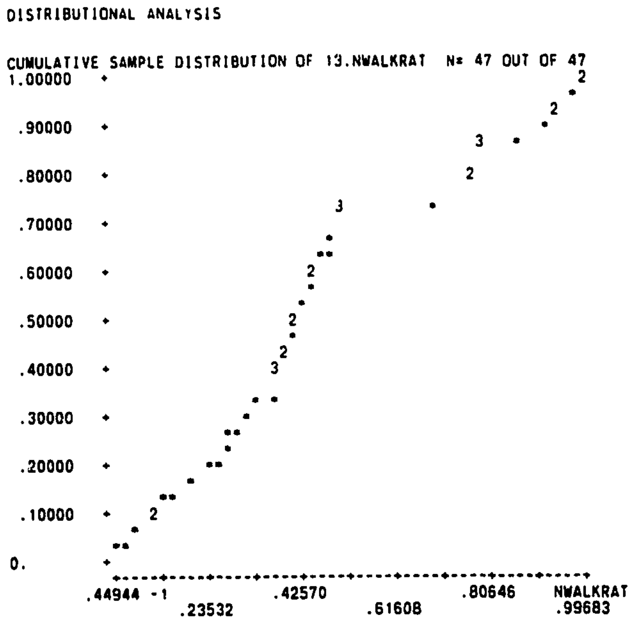
Cluster 3 Wait Time Ratio Normal Probability Plot



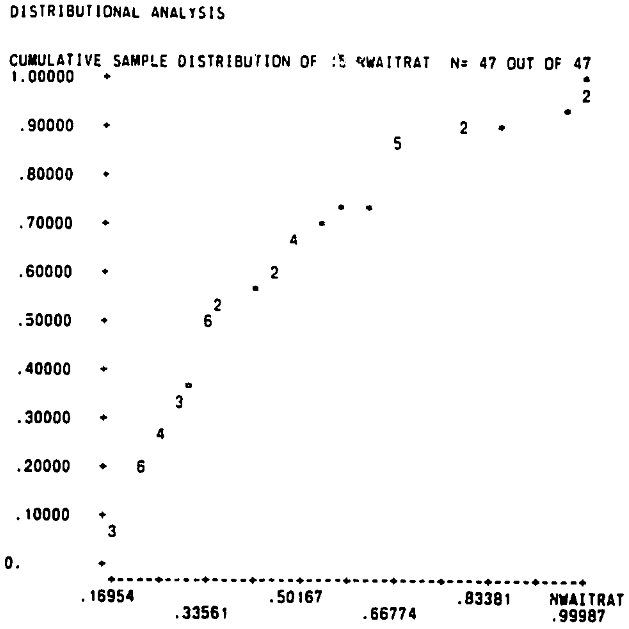
Cluster 3 Walk Time Difference Normal Probability Plot



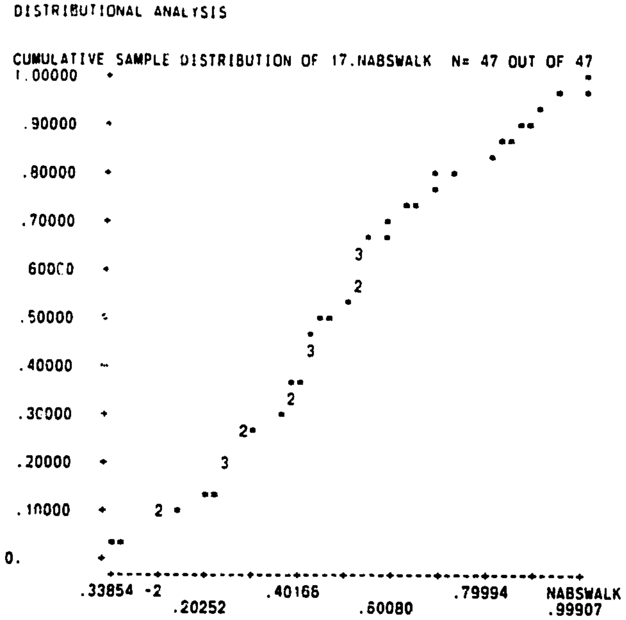
Cluster 3 Wait Time Difference Normal Probability Plot



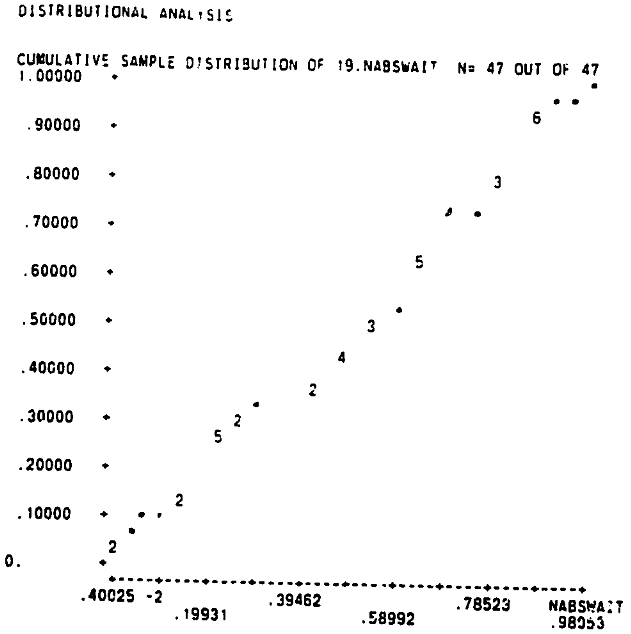
Cluster 4 Walk Time Ratio Normal Probability Plot



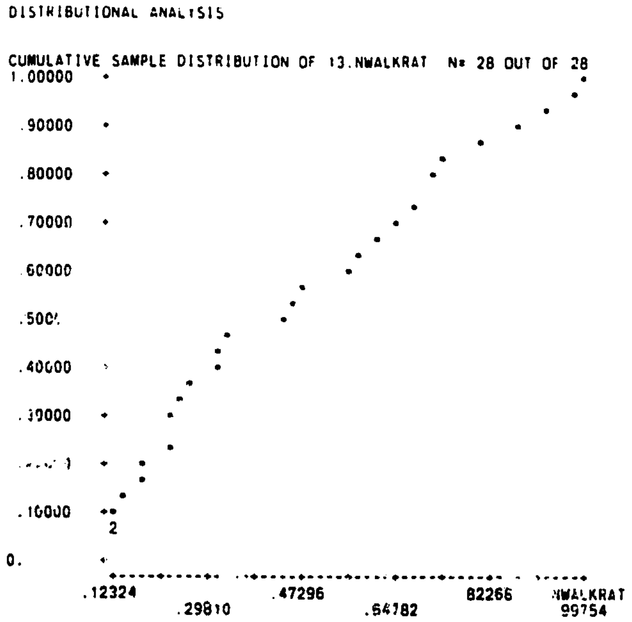
Cluster 4 Wait Time Ratio Normal Probability Plot



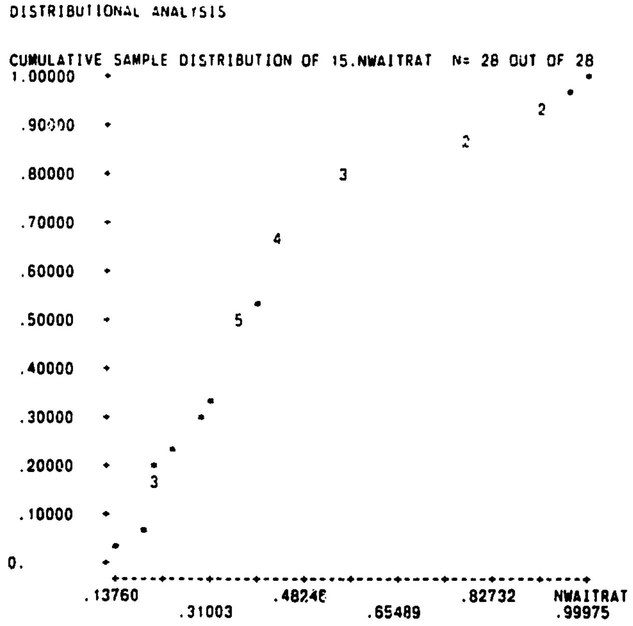
Cluster 4 Walk Time Difference Normal Probability Plot



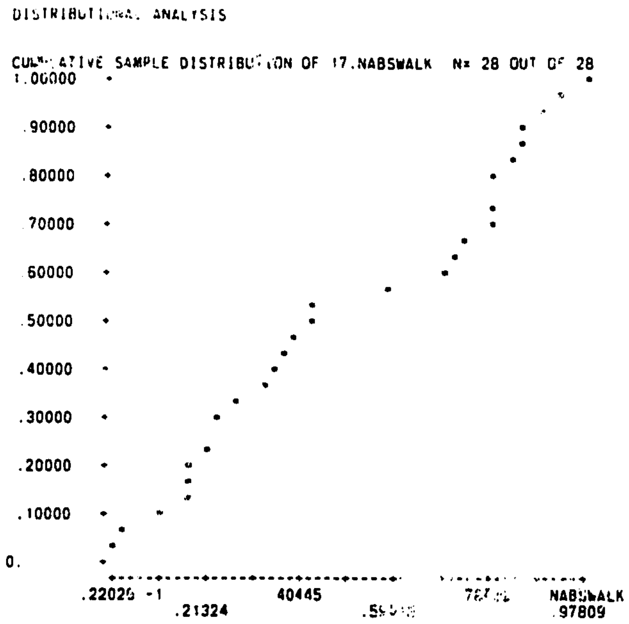
Cluster 4 Wait Time Difference Normal Probability Plot



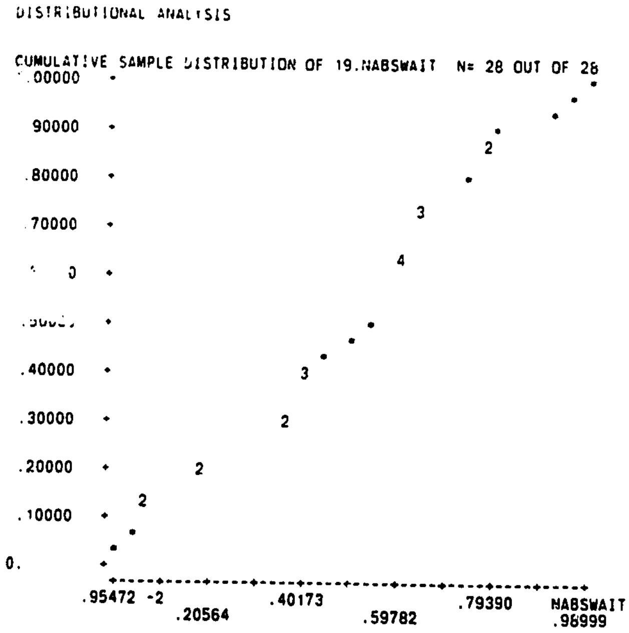
Cluster 5 Walk Time Ratio Normal Probability Plot



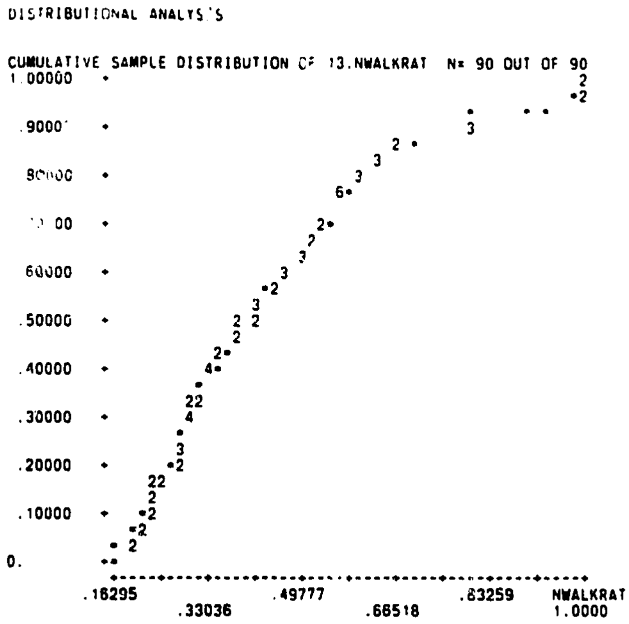
Cluster 5 Wait Time Ratio Normal Probability Plot



Cluster 5 Walk Time Difference Normal Probability Plot

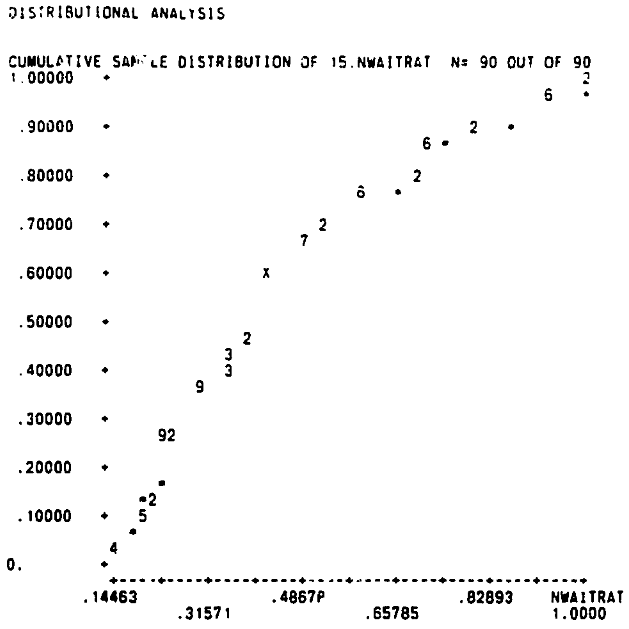


Cluster 5 Wait Time Difference Normal Probability Plot



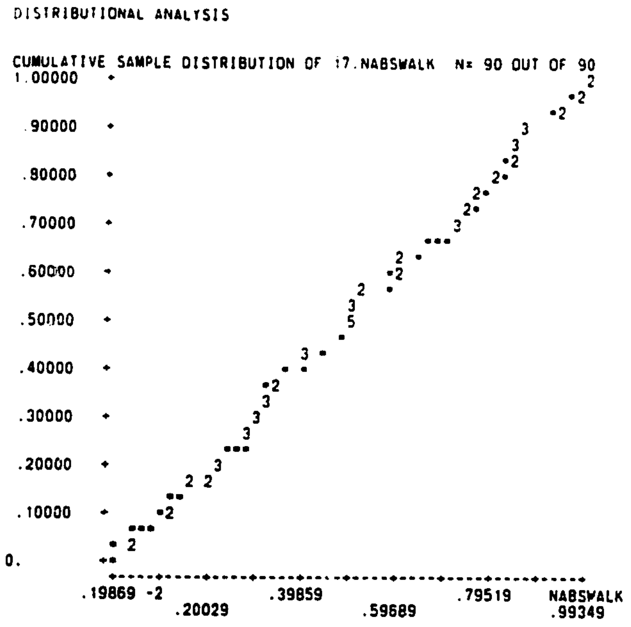
```
DISTRIBUTIONAL ANALYSIS

CUMULATIVE SAMPLE DISTRIBUTION OF 13.NWALKRAT  N= 90 OUT OF 90
```

Cluster 6 Walk Time Ratio Normal Probability Plot

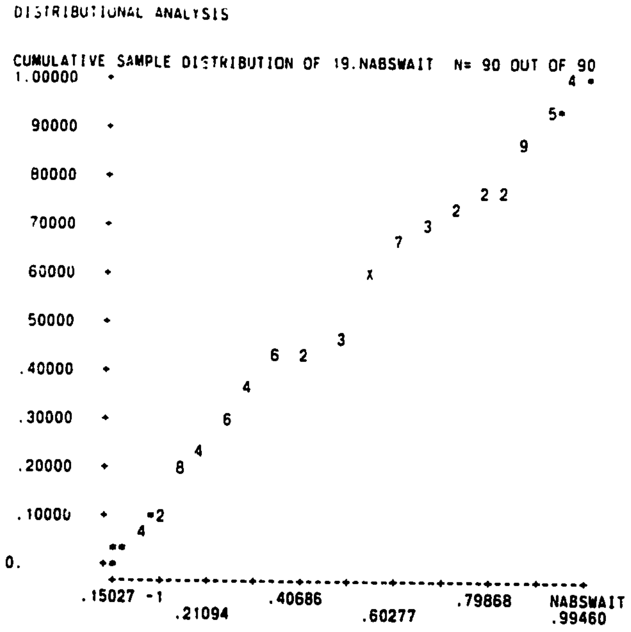


```
DISTRIBUTIONAL ANALYSIS

CUMULATIVE SAMPLE DISTRIBUTION OF 15.NWAITRAT  N= 90 OUT OF 90
```

Cluster 6 Wait Time Ratio Normal Probability Plot



Cluster 6 Walk Time Difference Normal Probability Plot



Cluster 6 Wait Time Difference Normal Probability Plot

Appendix F
F-Test and T-Test Results

WALK RATIO "F" VALUES

CLUSTER	1	2	3	4	5	6
1	*	1.162	1.141	1.997*	1.496	1.131
2		*	1.325	2.319**	1.738	1.314
3			*	1.750*	1.312	1.008
4				*	1.334	2.876**
5					*	1.323
6						*

WAIT RATIO "F" VALUES

CLUSTER	1	2	3	4	5	6
1	*	1.094	1.213	1.001	1.322	1.060
2		*	1.327	1.092	1.209	1.160
3			*	1.215	1.604	1.144
4				*	1.321	1.062
5					*	1.402
6						*

ABSOLUTE WALK TIME "F" VALUES

CLUSTER	1	2	3	4	5	6
1	*	1.015	1.092	1.092	1.181	1.059
2		*	1.109	1.108	1.164	1.044
3			*	1.000	1.290	1.157
4				*	1.290	1.157
5					*	1.115
6						*

ABSOLUTE WAIT TIME "F" VALUES

CLUSTER	1	2	3	4	5	6
1	*	1.340	1.392	1.262	1.468	1.109
2		*	1.039	1.061	1.096	1.208
3			*	1.102	1.055	1.255
4				*	1.163	1.138
5					*	1.324
6						*

ALL "F" VALUES ARE SIGNIFICANT AT A 5% LEVEL OF SIGNIFICANCE, EXCEPT FOR: (1)THOSE SIGNIFICANT AT A 1% LEVEL OF SIGNIFICANCE (INDICATED BY AN ASTERIX) AND (2)THOSE NOT SIGNIFICANT AT EITHER OF THESE SIGNIFICANCE LEVELS (INDICATED BY A DOUBLE ASTERIX).

Walk Ratio T-Statistics (alpha=0.8)

Clusters	1	2	3	4	5	6
1		-0.24	1.02	2.09	1.47	1.90
2			1.28	2.43	1.76	2.22
3				0.93	0.40	0.77
4					-0.55	-0.17
5						0.38
6						

Wait Ratio T-Statistics (alpha=0.8)

Clusters	1	2	3	4	5	6
1		0.31	-0.30	1.21	1.76	0.53
2			-0.60	1.04	1.68	0.24
3				1.39	1.87	0.80
4					0.61	-0.90
5						-1.59
6						

Walk Time Difference T-Statistics (alpha=0.8)

Clusters	1	2	3	4	5	6
1		0.98	-0.84	-1.93	-1.35	-2.45
2			-1.94	-3.25	-2.27	-3.94
3				-1.08	-0.66	-1.59
4					0.18	-0.49
5						-0.52
6						

Wait Time Difference T-Statistics (alpha=0.8)

Clusters	1	2	3	4	5	6
1		-0.22	-0.14	-1.70	-2.15	-1.01
2			0.05	-1.38	-1.90	-0.67
3				-1.27	-1.79	-0.62
4					-0.82	1.01
5						1.64
6						

* Shading indicates T-Statistics > 1.67

Appendix G
SPSS-X Regression Output

```
 7 Apr 90   SPSS-X RELEASE 3.0 FOR IBM MTS                                                               Page   1
17:11:32    University of Alberta

For                         University of Alberta                  License Number 30
This software is functional through January 31, 1991.

    1  0  data list file='gs74:data.dat' free
    2  0    /#id pwalk pwait awalk await compu income age gender job
    3  0     mode modefreq
    4  0  value labels
    5  0    income 03 '0000-4,999 $/yr' 04 '5,000-9,999 $/yr'
    6  0     05 '10,000-14,999 $/yr' 06 '15,000-19,999 $/yr'
    7  0     07 '20,000-24,999 $/yr' 08 '25,000-29,999 $/yr'
    8  0     09 '30,000-34,999 $/yr' 10 '35,000-39,999 $/yr'
    9  0     11 '40,000-44,999 $/yr' 12 '45,000-49,999 $/yr'
   10  0     13 '50,000-59,999 $/yr' 14 '60,000-69,999 $/yr'
   11  0     15 '70,000 $/yr and over'
   12  0    /age 1519 '15 to 19' 2024 '20 to 24' 2529 '25 to 29'
   13  0     3034 '30 to 34' 3539 '35 to 39' 4044 '40 to 44'
   14  0     4549 '45 to 49' 5054 '50 to 54' 5559 '55 to 59'
   15  0     6064 '60 to 64' 6569 '65 to 69'
   16  0    /gender 1 'male' 2 'female'
   17  0    /job 1 'most prestigous' 5 'least prestigous'
   18  0    /mode 0 'car drivers' 1 'public transport users'
   19  0    /modefreq 1 'less frequent users of usual mode'
   20  0     2 'usual mode used everyday'
   21  0  regression vars=pwalk to modefreq
   22  0    /statistics all
   23  0    /dependent=pwalk pwait
   24  0    /stepwise

There are 203384 bytes of memory available.
The largest contiguous area has 203032 bytes.

     4164 bytes of memory required for REGRESSION procedure.
        0 more bytes may be needed for Residuals plots.
```

```
                                   * * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Listwise Deletion of Missing Data

Equation Number 1    Dependent Variable..   PWALK

Beginning Block Number  1.  Method:  Stepwise

Variable(s) Entered on Step Number  1..    AWALK

Multiple R           .50342                                          Analysis of Variance
R Square             .25343       R Square Change     .25343                             DF      Sum of Squares      Mean Square
Adjusted R Square    .25202       F Change         179.91655         Regression           1          1060.76160       1060.76160
Standard Error      2.42814       Signif F Change     .0000          Residual           530          3124.80231          5.89585

                                                                     F =      179.91655       Signif F =  .0000

Condition number bounds:        1.000,      1.000

Var-Covar Matrix of Regression Coefficients (B)
Below Diagonal:  Covariance     Above:  Correlation

             AWALK

AWALK      3.209E-07

XTX Matrix

               AWALK  |    PWALK  |     AWAIT     COMPU    INCOME       AGE    GENDER       JOB      MODE  MODEFREQ

AWALK        1.00000  |  -.50342  |    .13704    .15740   -.00205    .03640    .00441   -.05642    .00814    .02247
----------------------+-----------+---------------------------------------------------------------------------------
PWALK         .50342  |   .74657  |    .04563   -.08194   -.03162    .04380    .05516    .01278   -.17633    .00520
----------------------+-----------+---------------------------------------------------------------------------------
AWAIT        -.13704  |   .04563  |    .98122   -.57458    .06265   -.00202   -.15047   -.05224   -.11174    .02340
COMPU        -.15740  |  -.08194  |   -.57458    .97523   -.14860   -.02685    .20411    .11173    .29918    .00971
INCOME        .00205  |  -.03162  |    .06265   -.14860   1.00000    .27111   -.59050   -.49044   -.20641    .03231
AGE          -.03640  |   .04380  |   -.00202   -.02685    .27111    .99868   -.16714   -.22958   -.04162    .11720
GENDER       -.00441  |   .05516  |   -.15047    .20411   -.59050   -.16714    .99998    .29167    .16064   -.00506
JOB           .05642  |   .01278  |   -.05224    .11173   -.49044   -.22958    .29167    .99682    .24934   -.06025
MODE         -.00814  |  -.17633  |   -.11174    .29918   -.20641   -.04162    .16064    .24934    .99993   -.06059
MODEFREQ     -.02247  |   .00520  |    .02340    .00971    .03231    .11720   -.00506   -.06025   -.06059    .99950
```

```
 7 Apr 90   SPSS-X RELEASE 3.0 FOR IBM MTS                                                                              Page   3
17:11:39    University of Alberta

                                  * * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Equation Number 1    Dependent Variable..   PWALK

------------------------------------------------------ Variables in the Equation ------------------------------------------------------

Variable              B         SE B    95% Confdnce Intrvl B       Beta     SE Beta   Correl Part Cor  Partial  Tolerance          T

AWALK           .007599  5.6652E-04       .006486     .008712    .503422     .037532  .503422  .503422  .503422   1.000000     13.413
(Constant)     1.627273     .183704      1.266396    1.988150                                                                  8.858

------ in -------            ------------------ Variables not in the Equation -------------------

Variable    Sig T            Variable       Beta In   Partial  Tolerance   Min Toler        T  Sig T

AWALK       .0000            AWAIT          .046508   .053318    .981220     .981220    1.228  .2200
(Constant)  .0000            COMPU        -.084019 -.096027    .975226     .975226   -2.219  .0269
                             INCOME       -.031624 -.036600    .999996     .999996    -.842  .4000
                             AGE            .043853   .050720    .998675     .998675    1.168  .2433
                             GENDER         .055162   .063842    .999981     .999981    1.471  .1418
                             JOB            .012825   .014819    .996817     .996817     .341  .7333
                             MODE         -.176337 -.204077    .999934     .999934   -4.795  .0000
                             MODEFREQ       .005198   .006014    .999495     .999495     .138  .8900

                                  * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Variable(s) Entered on Step Number  2..    MODE

Multiple R             .53341                                     Analysis of Variance
R Square               .28453        R Square Change      .03109                       DF      Sum of Squares      Mean Square
Adjusted R Square      .28182        F Change           22.98898    Regression            2          1190.90190        595.45095
Standard Error        2.37928        Signif F Change      .0000     Residual            529          2994.66201          5.66099

                                                                  F =       105.18501       Signif F =  .0000

Condition number bounds:       1.000,       4.000
```

```
 7 Apr 90   SPSS-X RELEASE 3.0 FOR IBM                                                                             Page    4
17:11:39    University of Alberta

                                        * * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Equation Number 1    Dependent Variable..    PWALK

Var-Covar Matrix of Regression Coefficients (B)
Below Diagonal:  Covariance    Above:  Correlation

                AWALK        MODE

AWALK       3.082E-07      .00814
MODE        1.057E-06      .05433

XTX Matrix

                AWALK        MODE |      PWALK |      AWAIT      COMPU     INCOME        AGE     GENDER        JOB   MODEFREQ

AWALK         1.00007      .00814 |    -.50199 |     .13795     .15496  -3.66E-04     .03674     .00310    -.05845     .02296
MODE           .00814     1.00007 |     .17634 |     .11175    -.29920     .20642     .04162    -.16066    -.24935     .06060
--------------------------------+------------+------------------------------------------------------------------------------
PWALK          .50199     -.17634 |     .71547 |     .02593    -.02918    -.06802     .03646     .08349     .05675    -.00549
--------------------------------+------------+------------------------------------------------------------------------------
AWAIT         -.13795     -.11175 |     .02593 |     .96873    -.54115     .03959    -.00667    -.13252    -.02438     .01663
COMPU         -.15496      .29920 |    -.02918 |    -.54115     .88571    -.08685    -.01439     .15605     .03713     .02784
INCOME       3.66E-04     -.20642 |    -.06802 |     .03959    -.08685     .95739     .26252    -.55734    -.43898     .01980
AGE           -.03674     -.04162 |     .03646 |    -.00667    -.01439     .26252     .99694    -.16046    -.21921     .11468
GENDER        -.00310      .16066 |     .08349 |    -.13252     .15605    -.55734    -.16046     .97417     .25161     .00467
JOB            .05845      .24935 |     .05675 |    -.02438     .03713    -.43898    -.21921     .25161     .93464    -.04515
MODEFREQ      -.02296     -.06060 |    -.00549 |     .01663     .02784     .01980      11468     .00467    -.04515     .99582


------------------------------------------------------ Variables in the Equation -----------------------------------------------

Variable               B        SE B     95% Confdnce Intrvl B       Beta     SE Beta   Correl Part Cor  Partial  Tolerance        T

AWALK            .007577  5.5514E-04     .006487     .008668     .501986    .036778  .503422  .501969  .510344    .999934   13.649
MODE           -1.120671     .233732   -1.579828    -.661514    -.176337    .036778 -.180425 -.176331 -.204077    .999934   -4.795
(Constant)      2.456684     .249653    1.966251    2.947118                                                                 9.840
```

```
                                    * * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Equation Number 1     Dependent Variable..    PWALK

------ In -------          ------------------ Variables not in the Equation ------------------

Variable      Sig T        Variable        Beta In  Partial  Tolerance  Min Toler          T  Sig T

AWALK         .0000        AWAIT           .026766  .031145    .968732    .968732       .716  .4743
MODE          .0000        COMPU          -.032946 -.035656    .885709    .885709      -.843  .3997
(Constant)    .0000        INCOME         -.071048 -.082187    .957389    .957330     -1.895  .0586
                           AGE             .036568  .043166    .996943    .996943       .993  .3213
                           GENDER          .085703  .100003    .974172    .974126      2.309  .0213
                           JOB             .050719  .069399    .934645    .934645      1.599  .1105
                           MODEFREQ       -.005512 -.006503    .995823    .995823      -.149  .8813


                                    * * * * * * * * * * * * * * * * * * * * * * * * * * * * *


Variable(s) Entered on Step Number  3..    GENDER


Multiple R            .54008                                        Analysis of Variance
R Square              .29168        R Square Change     .00716                          DF      Sum of Squares      Mean Square
Adjusted R Square     .28766        F Change           5.33370      Regression           3          1220.85054        406.95018
Standard Error       2.36960        Signif F Change     .0213       Residual           528          2964.71337          5.61499

                                                                    F =       72.47571       Signif F =  .0000

Condition number bounds:      1.027,      9.159


Var-Covar Matrix of Regression Coefficients (B)
Below Diagonal:  Covariance     Above:  Correlation

                AWALK        MODE     GENDER

AWALK      3.057E-07      .00753     .00314
MODE       9.822E-07      .05562    -.16065
GENDER     3.797E-07    -.00828      .04777
```

```
 7 Apr 90   SPSS-X RELEASE 3.0 FOR IBM MTS                                                                            Page   6
17:11:39    University of Alberta

                                        * * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Equation Number 1    Dependent Variable..   PWALK

XTX Matrix

                 AWALK       MODE      GENDER  |    PWALK  |    AWAIT      COMPU     INCOME        AGE        JOB   MODEFREQ

AWALK          1.00008     .00763     .00318  |  -.50225  |   .13837     .15446     .00141     .03725    -.05925     .02295
MODE            .00763    1.02656    -.16491  |   .19011  |   .08990    -.27347     .11451     .01516    -.20786     .06137
GENDER          .00318    -.16491    1.02651  |  -.08570  |   .13603    -.16018     .57211     .16471    -.25828    -.00480
------------------------------------------------+-----------+------------------------------------------------------------------
PWALK           .50225    -.19011     .08570  |   .70832  |   .03729    -.04255    -.02026     .05021     .03519    -.00589
------------------------------------------------+-----------+------------------------------------------------------------------
AWAIT          -.13837    -.08990    -.13603  |   .03729  |   .95070    -.51992    -.03623    -.02850     .00985     .01727
COMPU          -.15446     .27347     .16018  |  -.04255  |  -.51992     .86071     .00243     .01131    -.00317     .02709
INCOME         -.00141    -.11451    -.57211  |  -.02026  |  -.03623     .00243     .63853     .17072    -.29503     .02247
AGE            -.03725    -.01516    -.16471  |   .05021  |  -.02850     .01131     .17072     .97051    -.17776     .11545
JOB             .05925     .20786     .25828  |   .03519  |   .00985    -.00317    -.29503    -.17776     .86966    -.04635
MODEFREQ       -.02295    -.06137     .00480  |  -.00589  |   .01727     .02709     .02247     .11545    -.04635     .99580

---------------------------------------------------- Variables in the Equation ----------------------------------------------------

Variable               B          SE B     95% Confdnce Intrvl B       Beta      SE Beta   Correl Part Cor  Partial  Tolerance        T

AWALK            .007581   5.5288E-04      .006495       .008667    .502252     .036628  .503422  .502232  .512440    .999924   13.712
MODE           -1.208174      .235844    -1.671482      -.744867   -.190106     .037110 -.180425 -.187630 -.217598    .974126   -5.123
GENDER           .504750      .218556      .075405       .934094    .085703     .037109  .052941  .084589  .100003    .974172    2.309
(Constant)      1.685954      .416164      .868414      2.503494                                                                  4.051

------ in -------          ----------------- Variables not in the Equation ------------------

Variable     Sig T         Variable      Beta In   Partial  Tolerance   Min Toler        T   Sig T

AWALK        .0000         AWAIT         .039220   .045438    .950705     .950705    1.044   .2969
MODE         .0000         COMPU        -.049440  -.054500    .860713     .860713   -1.253   .2108
GENDER       .0213         INCOME       -.031723  -.030119    .638530     .638530    -.692   .4894
(Constant)   .0001         AGE           .051733   .060556    .970514     .948347    1.393   .1643
                           JOB           .040461   .044833    .869658     .869658    1.030   .3034
                           MODEFREQ     -.005915  -.007013    .995801     .970551    -.161   .8722

End Block Number   1   PIN =     .050 Limits reached.
```

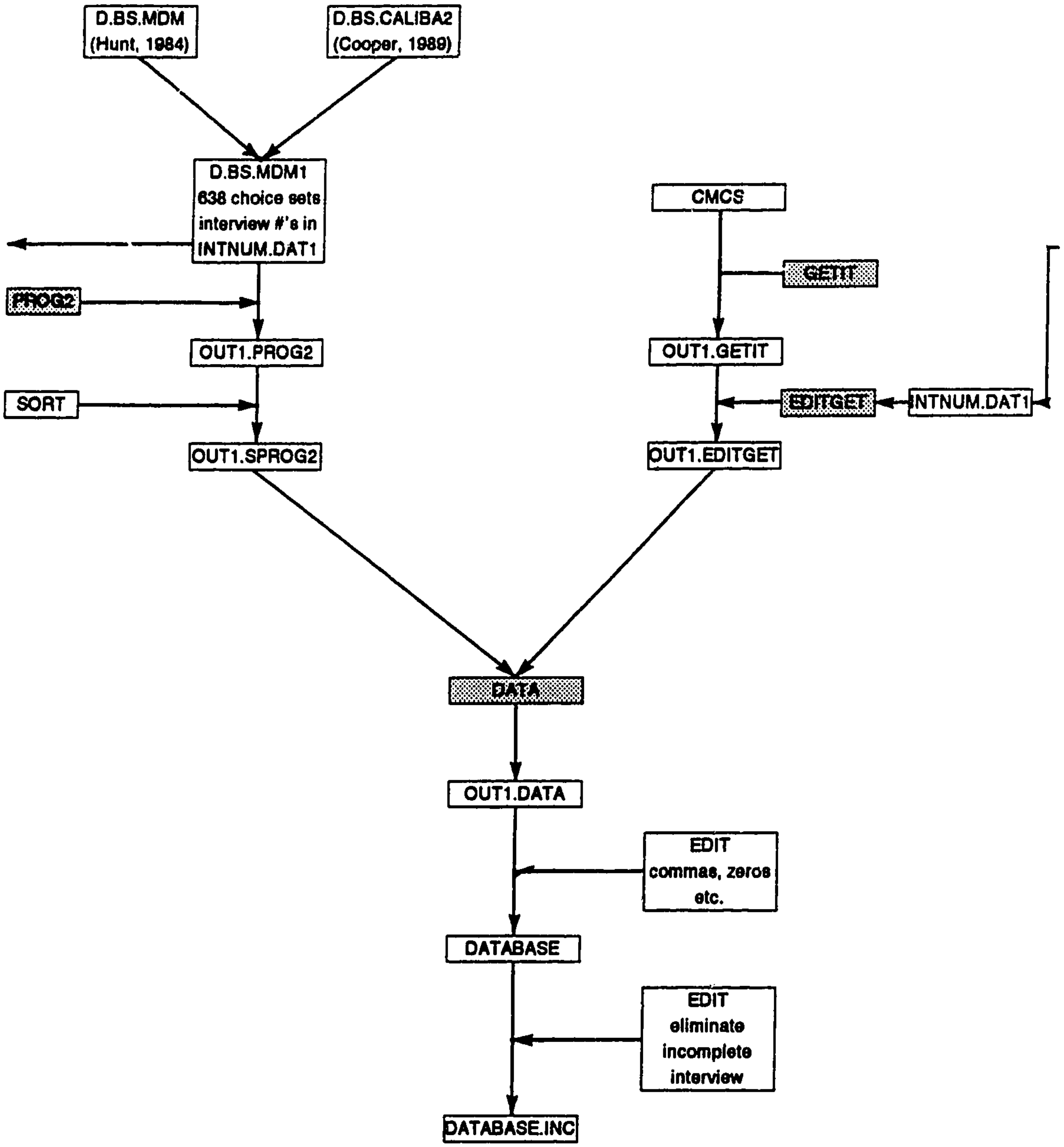
```
 7 Apr 90   SPSS-X RELEASE 3.0 FOR IBM MTS
17:11:39    University of Alberta                                                                          Page   7

                                  * * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Equation Number 1    Dependent Variable..   PWALK

                                                         Summary table
                                                         -------------

Step  MultR     Rsq  AdjRsq   F(Eqn)  SigF   RsqCh       FCh SigCh      Variable  BetaIn   Correl
   1  .5034   .2534   .2520  179.917  .000   .2534   179.917  .000  In:  AWALK      .5034    .5034
   2  .5334   .2845   .2818  105.185  .000   .0311    22.989  .000  In:  MODE      -.1763   -.1804
   3  .5401   .2917   .2877   72.476  .000   .0072     5.334  .021  In:  GENDER     .0857    .0529
```

```
                                  * * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Equation Number 2    Dependent Variable..   PWAIT

Beginning Block Number  1.  Method:  Stepwise

Variablo(s) Entered on Step Number  1..    MODE

Multiple R           .32540                                          Analysis of Variance
R Square             .10589        R Square Change    .10589                           DF      Sum of Squares      Mean Square
Adjusted R Square    .10420        F Change         62.76716         Regression         1           863.64064        863.64064
Standard Error      3.70937        Signif F Change    .0000          Residual         530          7292.50034         13.75943

                                                                     F =       62.76716       Signif F =  .0000

Condition number bounds:       1.000,       1.000

Var-Covar Matrix of Regression Coefficients (B)
Below Diagonal:  Covariance    Above:  Correlation

                 MODE

MODE           .13278

XTX Matrix

                 MODE |       PWAIT |      AWALK      AWAIT      COMPU     INCOME        AGE     GENDER        JOB   MODEFREQ

MODE          1.00000 |      .32540 |     .00814     .11063    -.30046     .20642     .04132    -.16068    -.24888     .06041
----------------------+-------------+-----------------------------------------------------------------------------------------
PWAIT         -.32540 |      .89411 |     .01424  7.75E-04      .01654  -8.41E-05     .01335     .02715     .01947     .00830
----------------------+-------------+-----------------------------------------------------------------------------------------
AWALK         -.00814 |      .01424 |     .99993    -.13794    -.15495   3.66E-04    -.03674    -.00310     .05844    -.02296
AWAIT         -.11063 |    7.75E-04 |    -.13794     .98776    -.51977     .03953    -.00160    -.13209    -.03244     .01980
COMPU          .30046 |      .01654 |    -.15495    -.51977     .90972    -.08690    -.00870     .15653     .02808     .03140
INCOME        -.20642 |   -8.41E-05 |   3.66E-04     .03953    -.08690     .95739     .26250    -.55734    -.43895     .01979
AGE           -.04132 |      .01335 |    -.03674    -.00160    -.00870     .26250     .99829    -.16034    -.22135     .11552
GENDER         .16068 |      .02715 |    -.00310    -.13209     .15653    -.55734    -.16034     .97418     .25143     .00474
JOB            .24888 |      .01947 |     .05844    -.03244     .02808    -.43895    -.22135     .25143     .93806    -.04649
MODEFREQ      -.06041 |      .00830 |    -.02296     .01980     .03140     .01979     .11552     .00474    -.04649     .99635
```

```
 7 Apr 90   SPSS-X RELEASE 3.0 FOR IBM MTS
17:11:39    University of Alberta                                                                              Page   9

                                    * * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Equation Number 2    Dependent Variable..  PWAIT

----------------------------------------------- Variables in the Equation ------------------------------------------------

Variable              B        SE B     95% Confdnce Intrvl B      Beta    SE Beta   Correl Part Cor  Partial  Tolerance        T

MODE          -2.886851     .364383    -3.602663   -2.171039   -.325405    .041073 -.325405 -.325405 -.325405   1.000000    -7.923
(Constant)     6.695035     .312385     6.081370    7.308701                                                                 21.432

------ in -------         ------------------ Variables not in the Equation ------------------

Variable     Sig T        Variable     Beta In  Partial  Tolerance  Min Toler        T  Sig T

MODE         .0000        AWALK        .014243  .015062    .999934    .999934     .346  .7291
(Constant)   .0000        AWAIT     7.843E-04  .000824    .987761    .987761     .019  .9849
                          COMPU        .018180  .018338    .909721    .909721     .422  .6733
                          INCOME   -8.786E-05 -.000091    .957389    .957389    -.002  .9983
                          AGE          .013376  .014134    .998292    .998292     .325  .7452
                          GENDER       .027868  .029089    .974182    .974182     .669  .5036
                          JOB          .020753  .021257    .938060    .938060     .489  .6250
                          MODEFREQ     .008332  .008795    .996351    .996351     .202  .8398

End Block Number   1   PIN =     .050 Limits reached.


                                    * * * * * * * * * * * * * * * * * * * * * * * * * * * *


                                                      Summary table
                                                      -------------

Step   MultR     Rsq  AdjRsq    F(Eqn)  SigF   RsqCh     FCh SigCh        Variable  BetaIn  Correl
  1    .3254   .1059   .1042    62.767  .000   .1059  62.767  .000   In:  MODE      -.3254  -.3254
```

Appendix H
Database Development Summary

FLOWCHART FOR CREATION OF RESEARCH DATABASE



Note: Programs written for database management have been highlighted by light shading.

Program name: PROG2

Source file name: S.PROG2

Object file name: O.PROG2

Program language: FORTRAN

Project for which it was developed: Perceptions

Author(s) name(s): Charlene Rohr

Date of last change: March 12, 1988

Description of purpose of program:

Determines, for automobile drivers, the walk distance and frequency for the public transit alternative the individual would choose as predicted by the TRAM logit model. For those individuals who choose public transit, the actual walk distance and frequency for their chosen public transit alternative are determined. The composite utility of all public transit alternatives, for each individual, is also calculated.

General Entry Points description: n/a

Subroutines stored in separate files: n/a

Support data files: n/a

Description of format of input files with examples included:

Input file, Unit 5, has the same format as that used for calibration of the TRAM model. See D.BS.MDM and D.BS.MDM1.

I/O Unit 5:

```
   3,3,81,0005,1,2529,09,
1,0317,0390,012,0,000,12,00,15,0122,0120,0,
2,0108,0260,001,0,000,13,00,03,0158,0425,0,
3,0108,0260,006,0,000,11,00,08,0131,0183,0,
   2,2,81,0010,2,2529,07,
1,0950,0180,055,0,000,22,00,10,0171,0191,0,
2,0948,0310,012,0,000,15,00,15,0122,0120,0,
   3,1,81,0014,2,2024,08,
1,1037,0250,009,0,000,12,00,05,0119,0503,0,
2,0969,0840,043,0,000,17,00,15,0132,0173,0,
3,1038,0470,049,0,000,10,00,03,0119,0503,0,
```

Description of format of ouput files with examples included:

Output format is (I4,',',I1,',',I4,',',I2,',',I4,',',I2,',',I1,',',F9.4,',') to correspond to the variables: NINT, NSEX, NAGE, NINC, MDISTO, MFREQT, MLRT, COMP.

I/O Unit 6:

```
 5,1,2529, 9, 260, 8,0,   -4.0859,
10,2,2529, 7, 310,15,0,   -5.6139,
14,2,2024, 8, 250, 5,0,   -4.8924,
```

Source file of Exec file: n/a

Description of MTS command to invoke program:

```
RUN O.SPROG2 5=D.BS.MDM1 6=OUT.PROG2
```

Extra comments:

none

Program name: DATA

Source file name: S.DATA

Object file name: O.DATA

Program language: FORTRAN

Project for which it was developed: Perceptions

Author(s) name(s): Charlene Rohr

Date of last change: Top of the morning, March 17, 1988

Description of purpose of program:

Creates data file for cluster analysis program by combining data from the Editget program and Prog2. The mode and mode frequency values are redefined as is described in Chapter 4.

General Entry Points description: n/a

Subroutines stored in separate files: n/a

Support data files: n/a

Description of format of input files with examples included:

Unit 4 reads in output from Editget program. For description see W.EDITGET.

I/O Unit 4:

```
0002 02 02 5 005 020
0005 05 01 5          005 003
0008 05 00 5          003r003r
0009 02 00 5 001 005
```

Unit 5 reads in the sorted output from the Prog2 program. To sort the output from Prog2 the following command was used:

```
r *sort par=sort=ch,a,1,4 input=out.prog2 output=out.sprog2
```

```
2,2,4549, 8,  150, 8,0,   -4.7290,
5,1,2529, 9,  260, 8,0,   -4.0859,
8,1,3034,12,  125,10,0,   -7.0110,
9,1,2529, 9,  105,10,0,   -3.8998,
```

Description of format of ouput files with examples included:

Output format is (3(I4),I5,I3,F7.2,I3,I5,4(I2)) to correspond to the variables INT1, PWALK, PWAIT, AWALK, FREQ, COMP, INC, AGE, GEND, JOB, MODE, MFREQ.

I/O Unit 6:

```
   2   5  20  150   8 -4.73   8 4549  2  5  0  3
   5   5   3  260   8 -4.09   9 2529  1  5  1  4
   8   3   3  125  10 -7.01  12 3034  1  5  1  0
   9   1   5  105  10 -3.90   0 2529  1  5  0  0
```

Source file of Exec file: n/a

Description of MTS command to invoke program:

```
RUN O.DATA 4=OUT.EDITGET 5=OUT.SPROG2 6=OUT.DATA
```

Extra comments:

none

Program name: EDITGET

Source file name: S.EDITGET

Object file name: O.EDITGET

Program language: FORTRAN

Project for which it was developed: Perceptions

Author(s) name(s): Charlene Rohr

Date of last change: March 10, 1988

Description of purpose of program:

Edits Getit output so that only the information for the interviews being investigated are given as output.

General Entry Points description: n/a

Subroutines stored in separate files: n/a

Support data files: n/a

Description of format of input files with examples included:

Unit 5 reads intnum.dat files which contain the information lines, sorted by interview number (increasing order), from D.BS.MDM files.

To create the intnum.dat files the following commands were used:

```
#edit D.BS.MDM
:col 1 3
:s@a/f '  '
:copy /s to file=-a
:mts
#r *sort par=sort=ch,a,11,14 input=-a output=intnum.dat
```

I/O Unit 5:

```
3,0,81,0002,2,4549,08,
3,3,81,0005,1,2529,09,
3,1,81,0008,1,3034,12,
3,0,81,0009,1,2529,09,
```

Unit 7 reads in the ouput from Getit. At present, the program is set up to read a maximum line length of 48 characters.

```
5 1707
  4      2      2      2      1
0.01A  0.07A  0.19   0.08A  0.15B
0001F02 02 06 5
0002 02 02 06 5
0003 05 05 01 5
0004 02 02 01 5
0005 05 05 01 5
0006 05 05 01 5
0007 02 02 01 5
0008 05 05 00 5
0009 02 02 00 5
```

Description of format of ouput files with examples included:
Output format is the same format as Getit program output.

I/O Unit 6:

```
2  2  2  6 5
5  5  5  1 5
8  5  5  0 5
9  2  2  0 5
```

Source file of Exec file: n/a

Description of MTS command to invoke program:

```
RUN O.EDITGET 5=INTNUM.DAT 7=OUT.GETIT 6=OUT.EDITGET
```

Extra comments:
none