

To the Graduate Council:

I am submitting herewith a thesis written by Roy Charles Longman entitled "An Evaluative Study of Geo-Demographic Database Systems." I have examined the final copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Geography.

Thomas L. Bell
Thomas L. Bell, Major Professor

We have read this thesis
and recommend its acceptance:

Ronald Forester

Accepted for the Council:

Lawrence Mink
Vice Chancellor
and Dean of the Graduate School

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Master's degree at the University of Tennessee, Knoxville, I agree that the Library shall make it available to borrowers under rules of the Library. Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgment of the source is made.

Permission for extensive quotation from or reproduction of this thesis may be granted by my major professor, or in his absence, by the Head of Interlibrary Services, when, in the opinion of either, the proposed use of the material is for scholarly purposes. Any copying or use of the material in this thesis for financial gain shall not be allowed without my written permission.

Signature Roy Charles Longman

Date November 21, 1990

An Evaluative Study of Geo-Demographic Database Systems

A Thesis

Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Roy Charles Longman

December 1990

ACKNOWLEDGMENTS

In the course of my thesis research, I received assistance from a number of scholars. In particular, I would like to thank Professor Thomas Bell for serving as my thesis director and to Professor Bruce Ralston and Professor Ronald Foresta for serving on my thesis committee. I am deeply indebted to them for their advice, guidance, criticism and encouragement throughout my thesis research. Whatever errors remain in my work, however, are solely my responsibility.

I would also like to thank Ms. Barbara O'Hare of the United Way of America for providing information about the United Way of America Geo-Demographic Database System. Without her assistance, it would not have been possible for me to complete my research.

ABSTRACT

The purpose of geo-demographic database systems is to develop a standardized classification which serves to identify homogeneous population groups according to similiar socio-economic and demographic characteristics. The census tract and ZIP code areal unit are normally utilized. Some applications of geo-demographic database systems are for trade area analysis, target marketing, and demographic studies. These systems represent an improvement over more classic urban structure models. They are based on sophisticated factor analytic and clustering techniques using a greater range of variables than classic urban structure models permitted, and are compatible with a variety of geographic scales. The thesis examines how these systems are developed, the methods used to develop these systems and how results are interpreted. A case study of the Knoxville MSA is undertaken at the census tract level.

Using a data set developed by the United Way of America, two research hypotheses are tested. Hypothesis one examines the sensitivity of group assignment based on the factor analytic and clustering techniques utilized. The results indicate that group assignments are sensitive to both the type of factor analytic and clustering techniques utilized. Hypothesis two examines the sensitivity of cluster group assignment to the input variables used.

To test hypothesis two, I developed an extended data set merging the United Way data with 23 variables selected from previous factorial ecology and quality-of-life studies.

The results indicate that the extended data set offers dramatic improvement in the amount of explained variance. Furthermore, cluster group assignment is more easily accomplished with this extended data set. Using standardized factor loadings, cluster means and standard deviations are calculated and the resultant clusters labeled.

Geo-demographic database system output based on the census tract areal unit is useful in providing general information about neighborhood composition. The results also indicate that the census tract is too heterogeneous to accurately segment the population into homogeneous groups. In the Knoxville MSA, some significant population groups existed within census tracts whose characteristics differ startlingly from the average tract composition and might, therefore, been overlooked. This aggregation problem may be solved if the data set used to develop the system could be based on forthcoming Census 90 data. Census 90 data will allow one to create smaller, more homogeneous areas than the census tract which should improve overall system usefulness.

TABLE OF CONTENTS

CHAPTER	PAGE
I. INTRODUCTION	1
Hypotheses and Research Methods.....	3
Hypothesis I: Sensitivity of Cluster Results To The Effects of Grouping.....	5
Hypothesis II: Sensitivity of Clusters to the Input Variables Used.....	7
Overview of the Thesis	11
II. INTRODUCTION TO GEO-DEMOGRAPHIC DATABASE SYSTEMS	13
Shortcomings Extant in Current Geo-Demographic Database Systems	16
The Relationship of Geo-Demographic Database Systems to Desktop Demographic Analysis Systems	20
III. LITERATURE REVIEW	25
Classic Urban Structure Models	25
Social Area Analysis	28
Modern Urban Structure Models - Factorial Ecology	31
Ecological Studies of Urban Structure Models Using Factor Analysis.....	32
Quality-of-Life-Studies Using Factor Analysis	36
IV. INTRODUCTION TO FACTOR ANALYSIS AND CLUSTER ANALYSIS	42
Factor Analytic Procedures	42
The Use of Cluster Analysis	50

V. DATA ANALYSES	58
Test 1. Comparing Commonly Used Factor Analytic Methods	59
Discussion of Table 5.9 Factor Loadings Within the Knoxville MSA	71
Test 2. Comparing Commonly Used Clustering Methods	75
Discussion of Hypothesis I	82
Test 3. Sensitivity of Clusters to the Input Variables Used	92
Discussion of Table 5.15 Factor Loadings Within the Knoxville MSA	98
Discussion of Hypothesis II	130
VI. CONCLUSION	133
Problems With Using Geo-Demographic Database Systems	137
Implications for Users	139
Areas of Future Research	141
REFERENCES	143
APPENDIXES	149
APPENDIX 1	150
APPENDIX 2	154
APPENDIX 3	155
VITA	157

LIST OF TABLES

TABLE	PAGE
3.1 SHEVKY AND BELL'S SOCIAL AREA ANALYSIS FACTORS	30
3.2 SUMMARY OF REES'S FACTOR ANALYSES	35
4.1 FORMULA FOR PRINCIPAL COMPONENTS ANALYSIS	45
4.2 FORMULA FOR PRINCIPAL AXIS FACTOR ANALYSIS	46
5.1 FACTOR ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS USING DEFAULT UNITED WAY SOLUTION OF PRINCIPAL COMPONENTS ANALYSIS PROMAX ROTATION.....	61
5.2 FACTOR ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS PRINCIPAL AXIS FACTOR ANALYSIS VARIMAX ROTATION.....	63
5.3 FACTOR ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS PRINCIPAL AXIS FACTOR ANALYSIS QUARTIMAX ROTATION.....	64
5.4 FACTOR ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS PRINCIPAL AXIS FACTOR ANALYSIS PROMAX ROTATION.....	65
5.5 FACTOR ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS PRINCIPAL COMPONENTS ANALYSIS VARIMAX ROTATION.....	66
5.6 FACTOR ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS PRINCIPAL COMPONENTS ANALYSIS QUARTIMAX ROTATION.....	67
5.7 FACTOR ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS PRINCIPAL COMPONENTS ANALYSIS PROMAX ROTATION.....	68
5.8 COEFFICIENT OF CONGRUENCE STATISTICS FOR KNOXVILLE MSA CENSUS TRACTS COMPARING FIVE FACTOR PRINCIPAL AXIS FACTOR ANALYSIS PROMAX ROTATION SOLUTION AND PRINCIPAL COMPONENTS ANALYSIS PROMAX ROTATION SOLUTION	70
5.9 FACTOR LOADINGS OF KNOXVILLE MSA CENSUS TRACTS USING PRINCIPAL COMPONENTS ANALYSIS PROMAX ROTATION BASED ON STANDARDIZED REGRESSION COEFFICIENTS.....	72
5.10 CLUSTER ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS USING PRINCIPAL COMPONENTS ANALYSIS PROMAX ROTATION BASED ON STANDARDIZED FACTOR SCORES.....	78
5.11 UNITED WAY'S DATA SET CLUSTER MEAN'S AND STANDARD DEVIATIONS FOR KNOXVILLE MSA CENSUS TRACTS.....	87

5.12	COMPARISON OF FACTOR ANALYTIC SOLUTIONS FOR KNOXVILLE MSA CENSUS TRACTS BETWEEN THE UNITED WAY'S DATA SET AND THE EXTENDED DATA SET USING PRINCIPAL COMPONENTS ANALYSIS PROMAX ROTATION.....	96
5.13	FACTOR ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS BASED UPON 23 VARIABLES EXTRACTED FROM PREVIOUS FACTORIAL ECOLOGY AND QUALITY-OF-LIFE STUDIES USING PRINCIPAL COMPONENTS ANALYSIS PROMAX ROTATION.....	97
5.14	FACTOR LOADINGS FOR KNOXVILLE MSA CENSUS TRACTS BASED UPON 23 VARIABLES EXTRACTED FROM PREVIOUS FACTORIAL ECOLOGY AND QUALITY-OF LIFE STUDIES BASED ON STANDARDIZED REGRESSION COEFFICIENTS.....	98
5.15	STANDARDIZED FACTOR LOADINGS FROM THE EXTENDED DATA SET FOR KNOXVILLE MSA CENSUS TRACTS USING PRINCIPAL COMPONENTS ANALYSIS PROMAX ROTATION BASED ON STANDARDIZED REGRESSION COEFFICIENTS.....	100
5.16	CLUSTER RESULTS FOR KNOXVILLE MSA CENSUS TRACTS COMPARING THE UNITED WAY'S DATA SET AND THE EXTENDED DATA SET.....	117
5.17	COMPARISON OF NEW CLUSTER ASSIGNMENTS BETWEEN THE UNITED WAY'S DATA SET AND THE EXTENDED DATA SET.....	118
5.18	EXTENDED DATA SET'S CLUSTER MEAN'S AND STANDARD DEVIATIONS FOR KNOXVILLE MSA CENSUS TRACTS.....	120

LIST OF FIGURES

FIGURE	PAGE
5.1 CLUSTER ASSIGNMENTS FOR ANDERSON COUNTY CENSUS TRACTS USING UNITED WAY DATA SET.....	79
5.2 CLUSTER ASSIGNMENTS FOR BLOUNT COUNTY CENSUS TRACTS USING UNITED WAY DATA SET.....	81
5.3 CLUSTER ASSIGNMENTS FOR KNOX COUNTY CENSUS TRACTS USING UNITED WAY DATA SET.....	83
5.4 CLUSTER ASSIGNMENTS FOR UNION COUNTY CENSUS TRACTS USING UNITED WAY DATA SET.....	85
5.5 CLUSTER ASSIGNMENTS FOR ANDERSON COUNTY CENSUS TRACTS USING EXTENDED DATA SET.....	109
5.6 CLUSTER ASSIGNMENTS FOR BLOUNT COUNTY CENSUS TRACTS USING EXTENDED DATA SET.....	111
5.7 CLUSTER ASSIGNMENTS FOR KNOX COUNTY CENSUS TRACTS USING EXTENDED DATA SET.....	113
5.8 CLUSTER ASSIGNMENTS FOR UNION COUNTY CENSUS TRACTS USING EXTENDED DATA SET.....	115

CHAPTER I

INTRODUCTION

The purpose of my thesis research is to examine the characteristics of geo-demographic database systems. Geo-demographic database systems are used to group neighborhoods and, ostensibly, the people who reside therein, into relatively homogeneous clusters. The neighborhood can be defined as a census tract, ZIP code, or other appropriate small areal unit. Geo-demographic database systems are being sold on a proprietary basis to a variety of commercial firms. They are used to target advertising for consumer products, determine the appropriate configuration of sales territories, and their classification schema may be correlated with a myriad of other phenomena which vary systematically by social standing and demographic composition. Retailers could use geo-demographic database systems to examine a particular market's demographic characteristics in order to determine what goods should be offered at a particular store. Location analysts might employ geo-demographic database systems to estimate trade area patronage.

In order to develop a geo-demographic database system, a set of variables is collected for the study area of interest.

These variables are felt to be important in identifying divergent demographic and socio-economic characteristics of the population at the appropriate areal unit - census tract or ZIP code for the more common for intra-urban study level or entire cities for an inter-urban comparison. Factor analysis is then performed on the data set. The table of factor loadings is used to label the dimension which each factor identifies. Then, the table of factor loadings is used as cluster analysis input. Cluster analysis assigns the areal unit (census tract, ZIP code area, or city) to the appropriate group based on the factor structure inherent in the socio-economic data (Weiss 1988; O'Hare 1988, Donnelly Marketing Information Services 1988). Currently there are four major geo-demographic systems commercially available. They are **CLUSTER PLUS** (Donnelly Marketing Information Services), **ACORN** (CACI Inc.), **VISION** (National Decision Systems), and **PRIZM** (Claritas Corporation). The number of clusters or neighborhood types produced by each system varies between 40 and 48 (Curry 1989).

Geo-demographic database system usage is not limited to the private sector. The United Way of America, a non-profit charitable organization, has developed its own geo-demographic database system to classify entire cities according to their estimated charitable contributions. The United Way geo-demographic database system uses both factor and cluster analysis to classify American cities by characteristics hypothesized to affect giving patterns and fundraising effectiveness (O'Hare 1988).

Geo-demographic database system output is one of many sources used in the more comprehensive desktop demographic analysis systems (Francesse 1990). Users of such desktop systems import data from many different sources, analyze them, and output them in the form of maps, charts, or reports. At present, there are five companies that are marketing desktop demographic analysis systems: **CONQUEST** (R. R. Donnelly Marketing Information Systems), **MARKET AMERICA** (CACI, Inc.), **PRIME LOCATION** (National Planning Data Corporation), **INFOMARK** (National Decision Systems) and **COMPASS** (Claritas Corporation) (Kirchner and Thomas 1989).

Hypotheses and Research Methods

It is beyond the scope of this research to examine the sophisticated desktop demographic systems currently available in any detail. Instead, I will concentrate solely on examining geo-demographic database systems which are a vital component of these linked desktop systems.

There does not seem to be any empirical research into the topic of geo-demographic database systems in any academic discipline. I find this surprising considering that important decisions are made by firms who put their trust in the integrity of this information. Perhaps because of the proprietary nature of these commercially available systems, there have not been any published reports which explain what data values are used, how these systems operate, or any empirical verification of their results.

I feel that there needs to be an empirical examination of these systems and this is the rationale underlying the present research.

In order to better understand how geo-demographic database systems function, I attempted to replicate an existing system. I contacted all four of the commercial vendors and the United Way of America in order to obtain information about their systems. Specifically, I asked them which variables were used to create their data set, and which factor and cluster analysis techniques were employed. All four of the commercial geo-demographic database system vendors were unwilling to release the detailed information needed to replicate their systems.

Fortunately, the United Way of America has graciously released information about the variables and methodology used in the creation of their geo-demographic database system (O'Hare 1986; O'Hare 1990) ¹. Even though the United Way's system is developed at the inter-urban level, the variables and research methodology can be applied an intra-urban study.

Whether the variables and procedures employed by the United Way are in any way comparable to the commercial systems cannot be determined. One of the persons responsible for the development of the United Way system used to work for one of the commercial vendors. It might be assumed, that the approaches are similiar even though the systems serve different functions. It is known that all of the five systems employ factor analysis and cluster analysis analytical tools.

1. I wish to thank Barbara O'Hare of the United Way for providing these data and the information about factor and clustering techniques employed.

Using the United Way's variables (Appendix 1) and research methodology, I developed an intra-urban geo-demographic database system for the Knoxville MSA at the census tract level. In 1980, the Knoxville MSA consisted of four counties: Anderson, Blount, Knox and Union. The total population of the MSA in 1980 was 476,517 (U.S. Bureau of the Census 1983b). This area was selected because I am familiar with Knoxville's population distribution and characteristics. I will, therefore, be able to examine the clusters produced to determine if they make intuitive sense.

I will also discuss how variables in a geo-demographic database system might be chosen, which factor and cluster analysis techniques could be employed and how the results might be interpreted. By performing all phases of system development, I am in a better position to identify problems that occur and gain an increased understanding of how these systems function. I also test the research hypotheses detailed in the two sections which follow.

Hypothesis I: Sensitivity of Cluster Results to the Effects of Grouping

I wish to examine changes, if any, in cluster assignments within a geo-demographic database classification system as a function of the type of factor and cluster techniques employed. First, I will determine if the factor results are sensitive to the analytical techniques employed. Second, I will assess whether the United Way's factor analytic procedures can be improved.

My hypothesis is that assignment of areal units to groups is highly sensitive to the particular factor and cluster analysis techniques used.

To test the sensitivity of geo-demographic database system classifications to the nuances of factor analysis, I will examine different factor analytic techniques, rotational methods, and changes in the number of identified factors. To evaluate factor analytic results, I will rely on three test statistics which will be detailed in Chapter IV. I will also rely on my ability to subjectively interpret differences in the results.

If each factor analytic technique indicates a rather different factor structure from those of its counterparts, I will accept my hypothesis that the type of procedure used is an important determinant of the resultant groupings. If each factor analytic technique produces a statistically similar structure, I will reject my hypothesis. There may be situations where two different factor analytic techniques produce solutions with similar amounts of explained variance, and the same number of factors, but with a different factor structure. In that case, I will use the coefficient of congruence to test the similarity of the factor loadings. If the coefficient of congruence statistics are statistically significant, I will reject my hypothesis that the factor analytic procedures used will significantly affect the factor results. If the coefficient of congruence statistics indicate dissimilarity of the results, I will accept my hypothesis.

I will compare the two factor analytic techniques and three rotational procedures discussed in Chapter IV to the United Way's factor analytic solution.

After examining the results, I will determine which factor analytic technique best represents the variance in the original data set. The factor loadings obtained from the best procedure will then be used as input to cluster analysis.

To examine changes in the number of clusters and cluster composition, I will rely on the test statistics which will be detailed in Chapter IV and my intuitive ability to interpret the results. Again, I am testing the sensitivity of cluster assignments to the analytical techniques employed. Second, I wish to determine if the United Way's cluster analysis method could be improved.

Hypothesis II: Sensitivity of Clusters to the Input Variables Used

After completing the research necessary to test whether grouping techniques employed in geo-demographic systems should make a difference in group formation, I will explore the importance of the specific variables used in the creation of geo-demographic systems. My hypothesis is that the system will better classify populations, if the variables used to create "modern" geo-demographic database systems are merged with variables used in classical urban ecological studies which are discussed in Chapter III. To test this hypothesis, I will create a new data set by merging the variables employed in the United Way system (Appendix 1) with an additional data set argued to be of importance in previous factorial ecology and quality-of-life studies (Appendix 2). My hypothesis is that this extended data set will increase the amount of explained variance, and develop more precise cluster groupings.

Based on this merged and extended data set, I will develop a geo-demographic database system for the Knoxville MSA at the census tract level. Using this extended data set, I will conduct factor analysis using the technique and rotation that worked best with the less extensive set of variables employed in the United Way's system. I will compare the results of my own geo-demographic system to the results obtained from the application of the United Way system. If the structure of the new data set warrants, I will to change the number of factors. Any change which I may be forced to make will not violate any factor analytic constraints.

Then, I will subject the factor loadings obtained from the extended variable set to cluster analysis. I will use the clustering method that best grouped the less extensive United Way data set. I will compare cluster groupings of the more comprehensive merged data set to those produced using the United Way data set in order to determine whether the extended set can be used to create more interpretable and meaningful clusters.

The procedures to be employed in the empirical portion of the thesis research may be summarized as follows:

Test Number I. - Comparing Commonly Used Factor Analytic Methods

Step 1.1 - Develop Database for the Knoxville MSA using the
United Way Geo-Demographic Database System's variables.

Step 1.2 - Replicate the United Way study at the intra-urban scale
using their variables and their factor analysis
methodology for the Knoxville MSA.

The United Way used Principal Components Analysis with promax rotation to develop their factor analytic solution.

Step 1.3 - Varying the number of factors, generate a number of factor analytic solutions. Principal Components Analysis and Principal Axis Factor Analysis with three different rotations (varimax, quartimax, promax) will be compared. The amount of explained variance, the Kaiser criterion, the $\pm .4$ factor loading cutoff, and subjective interpretation will be used to determine how many factors are needed to express the appropriate dimensionality of the data set.

Step 1.4 Compare the factor analytic results to determine which method best classifies this data set and compare it to United Way's factor solution (if they prove to be different). The best solution technique will be the one which maximizes explained variance without violating the aforementioned constraints.

Test Number II - Comparing commonly used cluster analytic methods

Step 2.1 - Using factor loadings determined from step 1.4, three different cluster analysis methods (Average, Single, Complete) will be performed and compared to the results of Ward's method used by the United Way. The scree test, changes in the number of clusters and subjective examination of the results will serve as goodness-of-fit criteria.

The process of determining which cluster analysis method is most appropriate, and the proper number of clusters is more subjective than the analysis of factor analytic solutions. Because the United Way's system is an inter-urban analysis, the number of clusters they derive cannot be used as a guideline for this intra-urban analysis.

Step 2.2 - Determine which cluster analysis method maximizes explained variance and best groups the data.

Test Number III - Determining the Sensitivity of Input Variables

Step 3.1 - Develop a new extended database based on United Way's variables and other variables determined to be of importance in previous factorial ecology/quality-of-life studies.

Step 3.2 - Develop a geo-demographic database system based on the extended database developed during step 3.1. Then, factor analyze this new database using the same factor analytic method and rotation created during step 1.4. If the structure of the variables warrant, change the number of factors.

Step 3.3 - Using factor loadings generated from step 3.2, use the clustering methodology from step 2.2 to subdivide the observational units into appropriate groups

Step 3.4 - Compare the clusters that were developed using the United Way's data set developed in step 2.2 to the

Overview of the Thesis

Chapter II provides an in-depth introduction into the commercial uses of geo-demographic database system output. The characteristics of each of the four major commercially available geo-demographic database systems are reviewed. Some problems that user's may encounter when using geo-demographic database system output will be discussed. The use of geo-demographic database system output as one of many input sources in desktop demographic analysis systems will be introduced and explained.

Chapter III reviews the literature from which many of the variables used in creating geo-demographic database systems are derived. This literature goes back to research conducted at the "Chicago school" of human ecology on internal differentiation within cities. It includes the contributions of urban scholars on city morphology, social area analysis and the more recent factorial ecologies. All of these literatures have contributed to our understanding of the spatial manifestations of the urbanization process within the modern Western city. All of these intellectual precursors, whether acknowledged or not, provide the basis for modern geo-demographic database systems.

Chapter IV provides an introduction to factor analysis and cluster analysis. Factor analysis and cluster analysis are the common techniques employed in these geo-demographic database systems.

Most users of these "user friendly" but opaque commercial systems are unfamiliar with the nuances of these data classification techniques which are used to group and classify observational units within geo-demographic database systems. That is, the techniques used to manipulate the input data are unknown to the user.

In Chapter IV, I will first provide an introduction to the statistical properties of factor analysis. I will then discuss the properties of the most commonly used factor analytic techniques and how the output from factor analysis is used in cluster analysis. Second, I will provide an introduction to the statistical properties of cluster analysis. I will review the most commonly used clustering procedures and discuss how each clustering technique is used to group data.

Chapter V will present and explain the results of the data analyses. Using the outline in Chapter I as a guide, I will assemble the necessary data and test the two major research hypotheses. At each step, I will present the results reporting relevant data as necessary.

Chapter VI will discuss my findings and summarize the thesis research. I will review the problems of developing geo-demographic database systems including factor and cluster analytic techniques, interpretation and usage of the results and limitations of current systems. I will identify other areas where I feel that more research should be undertaken.

CHAPTER II

INTRODUCTION TO GEO-DEMOGRAPHIC DATABASE SYSTEMS

There are four major commercial geo-demographic database systems available. In this chapter, I reviewed each of these systems and discussed their similarities and differences. Many of my sources for information about geo-demographic database systems were from reading promotional brochures and advertisements. Due to the vendors unwillingness to release information about their systems, I was not able to empirically verify the information that I discussed.

Probably the best known of the geo-demographic clustering systems is the Claritas Corporation's PRIZM system. In 1971, social scientist and computer programmer Jonathan Robbin created the Claritas Corporation to market his geo-demographic database system. Using 1970 census data, Robbin developed a computer program to transform census tract data into ZIP code format. Factor analysis was then performed on an undisclosed number of census variables at the ZIP code level of aggregation. From the factor score results, six factors were determined to be significant. They were social rank, mobility, ethnicity, family life style, housing composition and housing type. These six factors explained 87% of the original variance in the census-derived variables. The first three factors extracted by Robbin are, in fact, almost identical to those which have been discussed in social analysis studies.

The relevance of the association between geo-demographic clustering and social area analysis will be considered more extensively in Chapter III of the thesis (Shevky and Bell 1955).

After these six factors were extracted from the census-derived variables, cluster analysis was applied to the standardized factor scores associated with each ZIP code unit in order to produce to forty relatively homogeneous clusters (lifestyle segments) on a ZIP Quality or ZQ scale. These clusters vary from Blue Blood Estates (ZQ 1) to Public Assistance (ZQ 40) (Weiss 1988). ¹.

National Decision System's VISION geo-demographic system produces a 48 lifestyle clusters. Socio-economic, demographic, and housing characteristics at the neighborhood level to develop cluster rankings. These cluster rankings provide detailed lifestyle, purchasing patterns, consumption patterns and general information that can be used by market analysts to group and compare consumers (National Decision Systems 1987).

National Decision System's VISION system is based on using 117 socioeconomic, housing, and demographic variables measured for all 280,000 block groups to develop their 48 cluster system.

¹. For an in-depth view of the PRIZM system see Michael J Weiss' The Clustering of America. (New York; Harper and Row, 1988) This book discusses the development of the system, describes the characteristics of each cluster group, and maps the national distribution of each cluster by over 200 television viewing/market areas. The book also matches the ZIP Quality (ZQ) clusters with data on consumption, political and lifestyle behavior.

VISION data is one part of the VISION marketing network, which is divided into four components; Market Research; Market Analysis; Advertising/Media; Direct Marketing. The Market Research component provides information for a business to understand its existing and potential customers and in which types of products and services those customers might be interested.

The Market Analysis component uses VISION clusters for target marketing for test marketing, site selection, trade areas analysis and other target segmentation studies for any geographic areal unit --state, census tract, ZIP code, block groups or user defined areas. The Advertising/Media component provides information about consumer media behavior to understand where would be the best sources to advertise a particular product or service. The Direct Marketing component provides mailing information for a company's products based on VISION clusters including ZIP code areas, carrier routes, subscription lists, all encoded by the VISION cluster (National Decision Systems 1989).

R. R. Donnelly's CLUSTERPLUS system was created by analyzing all 1600 census-collected variables for 250,000 block groups. Of these initial 1600 variables, 64 were used in a factor analysis. The results indicated six factors: income, educational level, family lifestyle, type and age of housing units, population density, and ethnicity which describe the residents of census tracts, ZIP codes, and block groups. Then, it appears that cluster analysis is performed. The results of the cluster analysis indicated that 47 groups are meaningfully interpretable. Donnelly conducts yearly updates of their system using its DQI² residential database.

This database contains 200 behavioral and demographic variables for 70 million households. If there are significant changes, then the observational units can be reassigned to new clusters. The DQI² database also maintains name, mailing address, and phone number information on the households. By using either addresses or phone numbers, a consumer can be located and then assigned to the appropriate cluster. The system can also provide extremely detailed customer profile information which can be used to determine to what marketing information should be targeted. Such household specificity is not available with any other system (Donnelly Marketing Information Services 1988).

Shortcomings Extant in Current Geo-Demographic Database Systems

Geo-demographic database systems are useful for supplying general socioeconomic data about urban population at a small areal unit level. There are, however, at least two shortcomings with these systems.

The first shortcoming is with the method for determining clusters which underlies these systems. Geo-demographic database systems develop their neighborhood clusters by subjecting a set of variables to factor analysis and subsequently to cluster analysis. There are many different factor analysis methods available and rotational schemes which can cause significant differences in the resulting neighborhood clusters. These differences, if ignored, potentially can create serious interpretation problems (Weiss 1970). Similarly, there are a number of cluster analysis techniques that can be employed (Weiss 1971).

Given the problems inherent in using factor analysis and cluster analysis, the classification systems developed to date may have significant differences. These differences, furthermore, may be ignored by users of these packaged database target marketing tools only at considerable peril to their market strategies. User-friendly systems, in which the methods used to classify areas are opaque to the user, can easily lead to misleading or misinterpreted results if great caution is not exercised.

The second shortcoming is the ability of these geo-demographic database systems to provide generalized clustered rankings based on homogeneous study areas using ZIP code or census tract areal units. Most marketers use ZIP code data because they can directly use it for target marketing and direct mail advertising. It is common to use census tract data converted to ZIP code areas. But, the current five digit ZIP code is not a static areal unit (Stebbins 1987). Currently, ZIP codes areas are not required to conform to existing municipal boundaries. It is possible, therefore, for two municipalities to share the same ZIP code. There is, therefore, no way to ensure that the ZIP code areas are homogeneous (Pols 1987). In Weiss's book, The Clustering of America, he presents a map of ZIP code 10019 in Manhattan consisting of 30,000 people is presented (Weiss 1988). Eight distinct clusters or "lifestyle segments" are detected by Claritas' PRIZM system within that ZIP code. These include ZQ clusters as disparate in their characteristics as "Urban Gold Coast", "Single City Blues", "Bohemian Mix", "Gray Power," "Hispanic Mix", "Downtown Dixie-Style", "Public Assistance", and "New Melting Pot".

The overall ZIP code area is classified as 'Bohemian Mix', which completely ignores the other seven cluster types which are present in the ZIP code area.

Using Fairfax County, Virginia as the study area, Morsch and Shaw (1987) researched the applicability of the ZIP+2 (7 digit) system. The purpose of their research was to improve the performance of Multiple Regression Models (MRA) used to estimate property values. From a statistical standpoint, these models operate well. They are not, however, calibrated to take into consideration neighborhood characteristics in determining property value. Neighborhood characteristics are certainly an important determinant of property value and their impact must be decided by local appraisers. The purpose of developing the ZIP+2 system was to improve the MRA model's performance. This more detailed level of spatial resolution should correct for the deficiency of using current ZIP code areas for neighborhood delineation. The results proved their hypotheses: the ZIP+2 system delineated more homogeneous areas and could better be used to indicate neighborhood characteristics. This is not surprising, considering that smaller areal units are examined. But the result is useful in light of this current research. It would be beneficial to use a more finely detailed areal unit than the currently accepted ZIP code as the appropriate unit of aggregation if these geo-demographic database systems are to be of greater utility.

Some researchers feel that census tracts are a better unit of analysis than the ZIP code area (Stevenson 1989; Edmonson 1989).

Generally speaking, census tracts cover smaller areal units, but their intra-tract homogeneity is also questionable. Census tracts are small statistical areas within MSA's. They usually have between 2,500 to 8,000 people and are considered homogeneous at the time the census is taken (U.S. Bureau of the Census 1983a). However, there are problems with using census tracts as the area of analysis. First, census tracts are only delimited for areas within MSA's, thus excluding all rural areas within the nation. Second, their size may vary to the point that they do not contain a consistent degree of intra-tract homogeneity. Third, changes in demographic composition of residents between census collection periods may cause changes in the composition of the population (Morsch and Shaw 1987).

Because of the problems associated with both ZIP code and census tract areal units, it is certainly likely that the cluster composition may not be based on homogeneous areas. Without homogeneous areal units, the results of any geo-demographic database system should be interpreted with caution. It appears that the ZIP+4 system may be used as a data collection unit for the first time with the 1990 census. If this happens, it will improve the use of ZIP codes as an appropriate areal unit for geo-demographic analysis while simultaneously increasing the data handling and storage requirements necessary to operate these systems. Depending upon the region, the ZIP+4 system will create areas consisting of between four and eighty households. (Riche 1990).

The Relationship of Geo-Demographic Database Systems to Desktop Demographic Analysis Systems

The results from geo-demographic database systems are useful in providing general information about neighborhood composition. However, this information is not, in itself, sufficient to be used for marketing purposes. Geo-demographic database system output is one of many data sources used in desktop demographic analysis systems (Francese 1990). Desktop demographic systems allow users to import data from many different sources, analyze them, and then output them in chart, map, and/or report form. These desktop systems are especially beneficial for target marketing efforts because data can be sorted at either the census tract level or ZIP code level. By developing these systems for personal rather than mainframe computers, the amount of previous computer experience necessary for operators is decreased. The cost of entry-level desktop demographic systems varies between \$9,000 to \$14,000. At present, five companies are marketing such systems: **CONQUEST** (R. R. Donnelly Marketing Information Systems), **MARKET AMERICA** (CACI Inc.), **INFOMARK**, (National Decision Systems) **COMPASS** (Claritas Corporation) and **PRIME LOCATION** (National Planning Data Corporation) (Kirchner and Thomas 1989). Of these companies, Claritas, National Decision Systems and Donnelly Marketing Information Service are viewed as becoming the leaders in the consumer information field in the 1990s (Schwartz 1990).

The Claritas Corporation and National Planning Data Corporation are subsidiaries of VNU, a Dutch owned publishing and information company. Its United States headquarters is in New York City and its target market is the financial services industry. National Decision Systems is a subsidiary of the Equifax Corporation. The Equifax Corporation targets insurance companies and banking institutions providing information pertaining to consumer credit histories, as well as general consumer information (Schwartz 1990). Donnelly Marketing Information Services is one of 26 divisions of Dun and Bradstreet. Dun and Bradstreet is marketing data on supermarket product movement, direct mail demographics, and television viewership (Mason 1990).

The advent of desktop demographic systems combined with 1990 census data is signaling a change in the marketing industry. No longer will marketing companies be able just to supply and analyze census data and demographic information. Now these data will be merged with other sources to provide a more comprehensive view of consumer behavior and a better indicator of marketing performance (Schwartz 1990).

Claritas' COMPASS system provides access to demographic data, the PRIZM cluster data described previously, retail sales potential data, daytime population data, shopping center information data, syndicated media and purchase behavior data. In addition, the COMPASS system has the ability to include user supplied data at various geographic levels: state, cities, ZIP codes, counties, or census tracts. The COMPASS system contains information for most geographic areas of the United States (Crispell 1989).

National Decision Systems' **INFOMARK** system is based on using the **VISION** geo-demographic database system to analyze and compare consumers in the top 50 markets in the United States (Toomey 1990). It contains information about consumer expenditures, financial and consumer demand reports, and demographic characteristics. This system is extremely flexible in determining target marketing areas because it can use either county, census tract, zip code, or user-defined units (National Decision Systems 1990). One example of this system's use was by Popeye's Famous Fried Chicken in their buyout of rival Church's Chicken. Because of the large number of Church Chicken restaurants (1200), it was impossible for Popeye's management personnel to visit each location. The **INFOMARK** system was, therefore, used to plot the location of Church's sites relative to the location of existing Popeye's restaurants. In this way, the company could better determine which former Church's sites would be converted and which would be demolished (PC Week 1987). Another use of the **INFOMARK** system was by the American Savings Bank of New York. This firm is using the system to analyze the demographics of consumers of financial services. The system helped the firm to determine who their potential clients are, how they could be contacted, and what were their special financial needs (Shoultz 1988).

R.R. Donnelly Marketing Information Systems' **CONQUEST** system is currently considered to be the most advanced desktop system available. This system has access to Simmons and MRI Marketing Data; the Shopping Center Database with information on 24,000 centers; information about 7.4 million businesses at the four digit SIC level;

household expenditure data for 21 different retailing establishments and for 250 products; census data; and the **CLUSTER PLUS** geo-demographic database system. The user is able to define market areas, adjust market area sizes in order to eliminate overlap, add symbols as necessary, and to conduct shopping center searches to determine where best to locate a proposed retail outlet. For national firms, this system can be used to identify states and regions with a high sales potential. For local firms, the system can provide trade area analysis down to the block group level. The J. Walter Thompson Agency is one of the first marketing firms to use the **CONQUEST** system. Currently, the company is working to select between 300 and 400 shopping centers distributed throughout the United States for a retail outlet catering to a specific market segment. The Thompson Agency feels that the **CONQUEST** system will enable them to determine which centers would be the best centers to target (Kalish 1989; Donnelly 1988; Wood 1988).

National Planning Data Corporation's **PRIME LOCATION** system has three components. **MAXOnline** is the system's database. It contains information on demographics, retail sales, health care statistics, Yellow Page data, and uses the Claritas Corporation's **PRIZM** geo-demographic database system. **MAXLink** is software designed so that the system can communicate with the mainframe system. **MapAnalyst** is used to create graphical output from the analyses. Data from National Planning Data Corporation is also used by competitors Claritas, National Decision Systems, and other firms (National Planning Data Corporation 1990; Toomey 1990).

CACI's **MARKET AMERICA** system consists of a database with over 1500 demographic and product data items.

Within this database, there are population data for 1970 and 1980 with estimates for 1990 and 2000. The system has socioeconomic data for 1980; current and estimated population by sex, age, and race, and ACORN geo-demographic cluster data. It contains expenditure data for consumer purchases of automobiles, beverages, media, apparel, and sporting goods. Analyses may be conducted using "normal" study area units: ZIP code, census tract, counties, MSA, state, as well as other user-defined areas.

The following chapter will review the literature from which many of the variables used in geo-demographic database systems are derived. This literature includes reviewing classic urban ecological models (concentric zone, sector, and multiple nuclei), social area analysis, factorial ecology and quality-of-life studies. Examples of each type of model are examined and discussed.

CHAPTER III

LITERATURE REVIEW

The development of geo-demographic database systems is based on the desire to understand urban populations at a detailed geographic scale. This information can then be used to determine if people in a particular area would be interested in buying a particular product.

There have been many previous attempts to develop models to predict urban structure, but these previous academic studies lacked direct market applicability. They focused more on the morphology of the urban environment than on the precise characteristics of the population. The three classic urban structure models -- concentric zone, sector, and multiple nuclei -- do, however, provide the conceptual and methodological base for the geo-demographic database systems used today.

Classic Urban Structure Models

The concentric zone model was developed by University of Chicago sociologist Ernest Burgess who compared city growth to growth in plant life (Park, Burgess and McKenzie 1925). The concept of plant succession was redefined. The pattern of changing land uses within cities was argued to evolve in an analogous manner to that of plant communities.

Within this context, growth and obsolescence provide the mechanism for change. This analogy implies that as each land use zone expands into the periphery, land rent changes will reflect the adaptation of these zones to their new function. One example of this dynamic adjustment process is the location of slum areas nearest the downtown in the "Zone of Transition". The structures in this area are among the oldest in the city, and land rents are the lowest. On the other hand, the zone of most desirable residential housing would be in the "Residential Zone" and "Commuting Zone." It appears that Burgess used segregation, labor force composition, mobility, and land values to delimit ecological zones in his concentric model using the City of Chicago as the study area (Park, Burgess, and McKenzie 1925; Bourne 1971). Two criticisms of the Burgess model are that it does not accurately explain the relationship between land use and transportation and that it does not explain the clustering of industrial land uses (Hartshorn 1980).

The purpose of land economist Homer Hoyt's sector model was to refine and improve Burgess's concentric zone model. Hoyt's sector model assumed that growth radiates outward from the central business district into the periphery following major transportation routes. Hoyt felt that the horizontal expansion of the central business district was beginning to decline in most cities because of skyscrapers and competition that the CBD faced from outlying business complexes. Furthermore, decreased delivery time and increased factory-to-owner purchases decreased the need for a large warehouse district in most cities.

It was no longer necessary to build manufacturing plants near residential areas because of the increased mobility provided by public transportation and the automobile. Commercial and retail activity began to leave the central business district, locating in areas of new residential growth (Hoyt 1939). Hoyt developed his model by collecting data on eight housing variables for 142 cities to determine residential land rents. His conclusion was that high class residential neighborhoods spread outward from the center of the city along major transportation routes towards higher ground, undeveloped waterfronts or wooded areas, and opposite the direction of smoke and fog. Further, high rent areas expand along lines of least resistance where there is no natural or manmade barrier (Hoyt 1939). A criticism of Hoyt's study was that he concentrated too heavily on housing costs and ignored the characteristics of the occupants (Berry and Horton 1970).

The development of Harris and Ullman's multiple nuclei model refined both the sector model and the concentric zone model. Both of these previous models were based on the assumption that intra-city land use focused around the central business district (i.e. the monocentric city). After studying and comparing these models, Harris and Ullman identified four reasons why land use does not focus around a single center (the CBD) but rather land uses accrete around multiple centers dispersed throughout the city.

1. Certain activities require specialized functions...
2. Certain like activities group together because they profit from collusion...
3. Certain urban activities are detrimental to each other...
4. Certain activities are unable to afford the high rents of most desirable areas (Harris and Ullman 1945, pp. 14-15)

One criticism that I have of the multiple nuclei model is that it appears to be developed more from intuition, than from the results of either systematic empirical study, or a process theory of urban structural change. Harris and Ullman discussed how St. Louis, Chicago, Salt Lake City, and London, have some characteristics of multiple nuclei cities, but they did not elaborate on their study results.

All three of these models have some ability to predict certain types of urban land use. However, these models are not very useful in accurately predicting future growth within cities (Harris and Ullman 1945). Furthermore, these models were used to predict general land use patterns, and do not provide detailed demographic data about the population on a small areal unit basis.

Social Area Analysis

To compensate for the weakness of the urban structure models, the social area (factor) analysis model was developed in the 1940s by sociologists Eshref Shevky and Marilyn Williams who were interested in studying Los Angeles's urban structure. At the time, there was no model that could be used to study urban structure of large cities.

The purposes of developing the social area analysis model were to increase overall understanding of urban agglomerations and to develop a framework for analyzing the social characteristics of urban population. Their hypothesis was that there were three social factors; social rank, urbanization, and segregation that would be useful in determining urban structure of large cities.

Using 1940 census data, Shevky and Williams collected three variables to represent the social rank factor, three variables to represent the urbanization factor, and one variable to represent the segregation factor for the 563 census tracts in Los Angeles. The mean scores for each of these three social factors were used to categorize each census tract into one of fifteen Social Areas. The results confirmed the hypothesis that these three factors were important for understanding urban spatial structure (Shevky and Williams 1949).

Making some minor change in data collection techniques, sociologists Eshref Shevky and Wendell Bell conducted a social area analysis for San Francisco collecting data for 243 census tracts for 1940 and 244 census tracts in 1950 (See Table 3.1). For each time period, a data set consisting of six variables argued to be related to the three social factors -- social rank, urbanization, and segregation --was collected at the census tract level. For each census tract, the authors developed indexes of each social factor for both decades. The total score from the three indexes was used to rank each census tract by its social placement. To test their results, it appears that Shevky and Bell used correlation analysis, but this is poorly explained in their discussion of the validation techniques employed. After comparing the predicted quality-of-life scores with the actual population distribution, it is apparent that their intra-urban classification method was effective in replicating the spatial structure of cities (Shevky and Bell 1955; Berry and Horton 1970).

TABLE 3.1
SHEVKY AND BELL'S SOCIAL AREA ANALYSIS FACTORS

Social Rank (Economic Status)	Urbanization (Family Status)	Segregation (Ethnic Status)
1. Occupation 2. Education	1.Type of dwelling unit 2.Women in labor force 3.Fertility	1. Ethnic Status

Source: Shevky and Bell, 1955

The major criticism of social area analysis has been the supposed lack of theoretical justification for using these three social factors (Berry 1972). Shevky and Bell explained the purpose of each factor, but they did not discuss what criteria were used to select the variables used to operationalize the factor model.

Modern Urban Structure Models - Factorial Ecology

Most subsequent research in intra-urban classification models has focused on the factor analysis technique without making theoretical assumptions as to what dimensions should be expected to emerge. This more inductive research design has been called factorial ecology. Factorial ecology studies use the factor analysis technique with a wider range of variables than the Social Area Analysis technique permitted. The purpose of factorial ecology models is to examine persistent social indicators without making any preconceived assumptions about the theoretical nature of their interrelationships. With the aid of mainframe computers, researchers were able to increase the number of variables that could be factor analyzed (Cutter 1985). To better explain how these studies are conducted, I will review three classic intra-urban factorial ecology studies: Robert Murdie's Factorial Ecology of Toronto, Berry and Tennant's Commercial Structure of Metropolitan Chicago, and Phillip Rees' Factorial Ecology of Chicago (Murdie 1969; Berry and Tennant, 1965; Rees, 1968).

Ecological Studies of Urban Structure Using Factor Analysis

The purpose of Murdie's study was to develop an urban ecological model which analyzed both the longitudinal and cross-sectional structural changes in metropolitan Toronto. His study tested Shevky and Bell's social area analysis categories -- economic status, family status, and ethnic status -- to determine if they were either sectorally or concentrically distributed. His hypothesis was that economic status would be sectorally distributed, family status would be concentrically distributed, and ethnic status would be distributed both concentrically and sectorally. Using 109 variables from the Canadian census, he developed three separate data sets for 1951, 1961, and changes between 1951 and 1961. Murdie then subjected each of these data sets to principal components analysis with varimax rotation. The results indicated six relevant factors: three taken from social area analysis -- economic status, family status, ethnic status -- and three others -- recent growth, people employed in service and clerical occupations, and household characteristics. His results confirmed his hypotheses that economic status was sectorally distributed, family status was concentrically distributed, and ethnic status exhibited characteristics of both.

The purpose of Berry and Tennant's study was to provide a framework to examine the suburban structure of the northeastern Illinois metropolitan area excluding the City of Chicago proper.

Using 50 variables from 147 municipalities with populations over 2,500, they performed a principal axis factor analysis with a varimax rotation. Seven relevant factors were produced -- size, social status, family structure, new suburbs, distance-density, housing vacancies and race. These factors accounted for 92% of the common variance in the original 50 variables (Berry and Tennant 1965).

The Rees study of Chicago represents an improvement of Berry and Tennant's study (Berry and Horton 1970). There were four research questions associated with this study. The first question was to determine what was the most appropriate scale to differentiate urban population. To address this question, community and municipality data were compared against data at the census tract level. The second question compared the resulting patterns using the socio-economic variables against classic urban ecological models. The third question was to determine the type of social areas and communities that are present in the metropolis and where they are located. The fourth question was how the distribution of these disparate communities affected residential choice.

Rees conducted four separate factorial ecology studies of Chicago. In all four studies, principal components factor analysis with varimax rotation was employed. The first three studies were based on 57 variables describing the metropolis, city, and suburbs respectively. The fourth study was based on 12 variables analyzed for the 1,324 census tracts of Chicago (Rees 1968). I will limit my discussion to Rees' conclusions and provide a summary of his factors (See Table 3.2) ¹.

As a result of Rees's research, four conclusions were reached about the Chicago metropolis. First, family status and social status were shown to be important contributors to population distribution within the metropolis. Second, social class variables were sectorally distributed, and family status variables were concentrically distributed. Third, the spatial patterns within cities may be a function of city size and stage of development. Fourth, minority group predictors emerged. Of these groups, the distribution of black population was most closely associated with the socio-economic status indicators. This indicator could thus be used to prove racial discrimination within the metropolis. (Rees 1968)

Even though Rees's is one of the most comprehensive factorial studies ever undertaken, I would like to discuss two criticisms. The first is Rees's attempt to divide the metropolitan area of Chicago into three different broad areal units; metropolis, city, and suburbs. I realize that he was testing the methodology of factor analysis, but by partitioning the metropolitan area, the identification of the relevant factors becomes more difficult.

Second, varimax rotation produces orthogonal or uncorrelated factors, which is not always realistic. Forcing race and socio-economic status to be uncorrelated for the City of Chicago in 1960 seems unrealistic.

1.. For a detailed review of Phillip Rees's research and a discussion of factorial ecology, see Chapter 10 of Brian Berry and Frank Horton's book Geographic Perspectives on Urban Systems. Englewood Cliffs, N.J.: Prentice Hall., 1970. This chapter is partially based on Phillip Rees' Master's Thesis in geography at the University of Chicago.. It is a very well written reference for understanding how to conduct factorial ecology studies.

TABLE 3.2
SUMMARY OF REES'S FACTOR ANALYSES

Factor	% Variance	Factor	% Variance	Factor	% Variance
Metropolis		City		Suburbs	
1.Socio-economic Status	17.8	Race and Resources	28.9	Socio-economic Status	18.9
2.Stage in the Life Cycle	14.2	Socio-economic Status	24.0	Stage in the Life Cycle	14.3
3.Race and Resources	13.1	Stage in the Life Cycle	12.0	Immigrant and Catholic status	8.7
4.Immigrant and Catholic Status	10.8	Population size and density	9.0	Race and Resources	7.1
5.Population size and density	7.5	Other nonwhite population;workers	5.1	Population size and density	7.0
6.Jewish and Russian Pop.	3.8	Jewish and Russian Pop.	4.0	Poles vs. Anglo Saxons	4.1
7.Housing built in 1940's; workers commute by car.	3.0	Housing built before 1940;workers commute by rail.	3.4	European ethnic groups	3.0
8.Irish and Swedish Pop.	2.6			Housing built in 1940s; workers commute by car.	2.8
9.Mobility	2.4			Jewish and Russian Population	2.4
10.Other nonwhites and Italians.	2.1			Housing Overcrowded; workers commute by rail.	2.2
Total Predicted Variance	<u>77.3</u>		<u>86.4</u>		<u>70.5</u>
Census Tracts					
1.Socio-economic Status	37.3				
2.Stage in the Life Cycle	22.3				
3.Race and Resources	19.3				
Total Predicted Variance	<u>78.9</u>				

Source: Rees' 1968

Considering how important race was in determining socio-economic status thirty years ago, perhaps an oblique rotation which would allow for a degree of intercorrelation among factors would have been more appropriate. Rees discussed how the distribution of black population in Chicago was strongly associated with socio-economic status. I feel, therefore, that the use of varimax rotation was inappropriate, in this case.

It is worth noting that the results of nearly all factorial ecology studies applied to United States cities demonstrates the robustness of sociological theory of social change caused by urbanization and confirm the relevance of Shvky and Bell's original three social area factors (Berry 1972; Cutter 1985).

Quality-of-Life Studies Using Factor Analysis

In the 1970s, the type of intra-urban analyses changed from the social area analysis/factorial ecology type to quality-of-life studies. The primary difference between these two classes of studies is the type of data utilized. Factorial ecology studies rely heavily on census data, while quality-of-life studies rely more on social data to develop quality-of-life indicators for specific communities (Dickinson et al. 1972). However, few quality-of-life studies have been completed at the intra-urban level.

To provide some background into the purposes that quality-of-life studies supposedly serve, I will review four studies: the Liu inter-urban study of United States cities and three intra-urban studies: the Dickinson et al. study of Gainesville, Florida, the Smith study of Tampa, Florida, and the Bederman study of Atlanta, Georgia (Liu 1976; Dickinson 1972; Smith 1973; and Bederman 1974).

The purpose of Liu's study was to develop general quality-of-life indicators based on data for all 243 SMSA's extant in 1970. To further refine his results, the SMSA's were divided into three size groupings -- large, medium, and small. For his study, 120 variables from five general components -- economic, political, environmental, health & education, and social -- were used. Three methods to analyze these quality-of-life measures were employed in this study -- component and factor analysis, a standardized additive method, and the adjusted additive method. The adjusted additive method was felt to be the best measure. The standardized data were converted into Z scores and the score's distance from the mean was used to develop a grading system. The points on the grading system ranged from 1 to 5 with 1 being the lowest and 5 being the highest. SMSA ratings for each of the five categories are calculated, and these ratings are totaled to calculate overall quality-of life (Liu 1976).

In the Dickinson et al. study of Gainesville, Florida, seventeen variables from five categories -- housing, home and family, crime, health, and poverty and welfare -- were collected at the enumeration district level.

Three alternative methods of classifying data were used in this study -- Z-scores, the McHarg map-overlay technique ¹. and principal components analysis. The results using each of these disparate measurement techniques were quite similiar. A transformation was performed on all the variables to assure normality of the data. The transformed scores were summed for each social indicator and divided by the number of variables in order to calculate the social indicator score. Then the scores from the five social indicators were summed and divided by the number of social indicators (5) to calculate the overall quality-of-life score for each individual census tract. Maps based on the distribution of these social indicators at the census tract level and the overall quality-of-life score for each census tract were plotted and were similiar.

Although, not an exact goodness-of-fit criterion, map comparison gives a qualitative indication of map similarity. To validate their study, correlation between the five social indicators and the quality-of-life indicator was conducted and all of the factors have at least a .83 correlation. From their results, it appears that the model provides a reasonable explanation of the spatial variation of the quality-of-life in this city (Dickinson et al. 1972).

The purpose of Smith's quality-of-life study was to develop a set of social indicators for the Metropolitan Development Agency for the city of Tampa, Florida.

¹. For a detailed explanation of the McHarg map overlay technique see McHarg, Ian L. Design With Nature. New York: The Natural History Press., 1969

In Smith's study, 47 variables from six categories measured within census tracts were used as input -- economic status, environment, health, education, social disorganization and participation, and equality. A principal components factor analysis with an oblique rotation was completed on the data set. The output from factor analysis produced four relevant factors: social problems, social-economic status, racial segregation and social deprivation. These four factors accounted for 47% of the variance in the data set. The spatial distribution of factor scores for each of these oblique factors was plotted by census tract. It was, therefore, possible to provide spatial information that the Metropolitan Planning Agency could use to target areas needing further study and assistance. Some similarities were seen in the distribution of the four factors, which is to be expected with oblique (correlated) rotational schemes. Because this study did not rely as heavily on census data, the results were different than what would be expected from factorial ecology/social area analysis. This is important, because Smith believes that social problems are more important than social area factors in measuring quality-of-life (Smith 1973).

In Bederman's quality-of-life study of Atlanta, eleven variables were collected for 112 census tracts. Instead of using factor analysis, Bederman used a weighted average technique to create five sub-groupings. First, average values for each variable in the city were used as the base value. Second, each variable was standardized. This standardized score is compared to the base value, and the difference in these values was recorded.

Third, this value was multiplied by 100. Fourth, the standardized variable values were divided by the number of variables used to distinguish the factor. Fifth, the factor scores were summed to determine the quality-of-life value for each census tract. Then, to better understand the effect of race, Bederman compared quality-of-life in both white and black areas. The mapped distribution obtained by these procedures appeared to confirm that quality-of-life in the city varies by race. Bederman's results confirmed his belief that race is an important factor in determining quality-of-life within Atlanta (Bederman 1974).

There are some problems associated with quality-of-life studies. The most important problems are the lack of a standardized data set and a standard method of analysis (Bederman 1974). Without a standardized data set, unique variables must be chosen for each city and this makes inter-city comparisons difficult. This is one problem that factorial ecology studies did not have. By relying on census data, similar studies in other communities could be replicated to test hypotheses using a commonly accepted methodological procedure (i.e. factor analysis). And yet, mathematically comparative factorial ecology would be impossible. To truly compare two or more cities with mathematical precision, the variance of the variables would have to be homoskedastic across all places: an unlikely situation indeed.

Another problem with quality-of-life studies is to objectively determine what the relevant social indicators should be (Smith 1973).

For example, in an all-white or nearly all-white community, race would not be a relevant factor in poverty, as it would be in a city with a large minority population, such as is the case in Atlanta (Bederman 1974).

In the 1980s, research into developing intra-urban models using either factorial ecology or the pursuit of improved quality-of-life studies has decreased. The factor analysis literature has changed from being concerned with the nuances of the technique *per se*, to interest in the technique within a context as a problem solving tool.

The following chapter will introduce and explain how factor and cluster analysis procedures are used to create geo-demographic database systems. I will begin by providing a brief introduction to the statistical properties of factor analysis. The most commonly used factor analytic procedures will then be reviewed and compared. Next, I will explain how factor analyzed data are used as cluster analysis input. I will discuss how cluster analysis is used to group observations in a data set. The most commonly used cluster analytic procedures will be reviewed and compared.

CHAPTER IV

INTRODUCTION TO FACTOR ANALYSIS AND CLUSTER ANALYSIS

As discussed in Chapter I and Chapter II, geo-demographic database systems rely upon factor analysis and cluster analysis techniques to create neighborhood clusters. There are a number of factoring and clustering techniques available to the researcher. Each technique has particular mathematical properties and assumptions underlying its use. Before making complex marketing decisions based on the results of a geo-demographic database system, the user should have some knowledge of how both of these techniques are used to classify data, which techniques are most commonly used by researchers, and how both techniques are used to create geo-demographic database systems. I will begin first by discussing factor analysis.

Factor Analytic Procedures

Factor analysis is a technique used to analyze, simplify and classify mathematically a set of variables. By extracting each variable's common variance, fractions of variables are partitioned among several factors (Rummel, 1967; Weiss, 1971). The number of identified factors reflects how many similar groups exist within the data set (Rummel, 1970).

Factor analysis is useful to parsimoniously reduce the size of a data set. If, for example, one recorded the values of 50 socio-economic variables for 100 cities, it would be difficult to interpret the results. The resultant matrix would contain 50,000 cells. However, by using factor analysis, the principal dimensions measured by the variables are classified into a smaller number of significant factors.

Another use of factor analysis is to transform raw data into standardized form. The standardized output can be used as input into other techniques such as regression and cluster analysis (Rummel, 1967; Taylor, 1977; Punj and Stewart, 1983). Factor analysis can be used with almost any data set, but there must be a large amount of variation among the variable values or the factor structure will not be clearly indicated (Rummel, 1970).

Factor analysis requires that the user develop a data matrix. Such a matrix usually consists of the observational units (e.g. census tracts, ZIP code zones, cities) as rows and the selected measurements on those observational units as columns. An intercorrelation matrix may be created algebraically by a transformation of the raw data matrix using the communality estimate value. The communality estimate value is the amount of correlation permitted between a variable and itself (i.e. the major diagonal elements of the intercorrelation matrix created by the raw data). One way that factor analysis methods differ is the method used to derive the communality estimate (Taylor, 1977).

Even though there are many types of factor analysis techniques available, the two most commonly used are principal axis factor analysis and principal components factor analysis.

The reliability of both of these techniques has been empirically verified in urban ecological studies, quality-of-life studies, and factorial ecology studies which were reviewed in Chapter II. The principal axis method was used, for example, in Berry and Tennant's study of metropolitan Chicago (Berry and Tennant, 1965). The principal components method, on the other hand, was used in Murdie's study of Toronto, Rees's study of Chicago, and Smith's study of Tampa (Murdie, 1969; Rees, 1968; Smith, 1973).

Principal components analysis involves a simple rescaling of the correlation matrix so that successive factors explain decreasing amounts of variance. The formula for principal components analysis is listed in Table 4.1 (page 45). In principal components analysis, the model is assumed to be closed and any unique variance that a variable might contribute to general explanation is ignored. This technique uses the communality estimate value of 1.0 which becomes the major diagonal of the intercorrelation matrix (Weiss, 1970). One advantage to principal components analysis is that the result is a mathematically determinate solution. A disadvantage is that if any unique variance occurs, the structure and the factors are more difficult to identify, because the unique variance will distort the structure (Taylor, 1977).

Common factor analysis (principal axis factor analysis) is used when researchers believe there is some unique variance in each variable. Unique variance may be due to measurement error or to a truly unique output of variables. Uncertainty about measurement reliability is quite common in social science research (Weiss, 1970). The formula for principal axis factor analysis is listed in Table 4.2 (page 46)

TABLE 4.1
FORMULA FOR PRINCIPAL COMPONENTS ANALYSIS

X_j - Variable Name
 S_i - Factor Name
 a_{ji} - The contribution of S_i to the common variance
 OF X_j
 p - Number of common factors

For each X_j

$$X_j = a_{j1}S_1 + a_{j2}S_2 + \dots a_{jp}S_p$$

Source: Rummel, 1970

TABLE 4.2
FORMULA FOR PRINCIPAL AXIS FACTOR ANALYSIS

X_j - Variable Name
 S_i - Factor Name
 a_{ji} - The contribution of S_i to the common variance of X_j
 p - Number of common factors
 S_{ju} - Factor contributing to the unique variance of X_j
 a_{ju} - The contribution of S_{ju} to the unique variance of X_j
 For each X_j

$$X_j = a_{j1}S_1 + a_{j2}S_2 + \dots a_{jp}S_p + a_{ju}S_{ju}$$

Source: Rummel, 1970

The principal axis technique is normally viewed as superior to principal components analysis because variables are permitted to have both common and unique variance. By allowing both types of variance, this method allows variables to have unique variance. The communality estimate is, therefore, less distorted. In principal axis factor analysis, the squared multiple correlation value is the most commonly used communality estimate value (Weiss, 1970). A disadvantage of principal axis factor analysis is that the results are not always mathematically determinate (Taylor, 1977).

Factor analysis produces an intercorrelation matrix consisting of factor loadings. All of the factor loadings which were created in the initial factor analysis step exist in unrotated form. The problem with forming factors based on unrotated factor loadings is that the some factors may be created from variables with high factor loadings and other factors created from variables with low factor loadings. The resulting factors may incorrectly identify the data structure (Rummel, 1967). Rotated factor solutions are preferred because they are easier to interpret, replicate, and are necessary to achieve simple structure. Simple structure attempts to ensure that only a few variables have high factor loadings and that these variables identify only one factor. (Weiss, 1971)

There are two commonly used factor rotation techniques: orthogonal and oblique. The difference between these techniques is that oblique rotations permit factors to be intercorrelated while orthogonal rotations do not (Rummel, 1967; Weiss, 1971).

Orthogonal rotations force factors to be uncorrelated. Orthogonal factor loadings are correlations and their squared values represent the amount of common variance in a variable which is explained by the factors. A high factor loading between a variable and a particular extracted factor indicates that it is important in factor formation. On the other hand, a low loading indicates that the variable is not important in formation of the particular factor. Each cluster of variables with similar variance is assigned to a factor. Two advantages to orthogonal rotations are that factor loadings can be used as correlation coefficients and the resultant factors are more easily interpreted. The disadvantage to orthogonal rotation is that orthogonality is imposed on the factors when, in fact, they may be intercorrelated. If these factors happen to be correlated, the identified factors may incorrectly identify the structure of the data set (Rummel, 1967; Weiss, 1971).

Among orthogonal rotations, varimax and quartimax are the two most commonly used to approximate simple structure. Varimax rotation maximizes the amount of each factor's explained variance. It is useful in simplifying factor description. An advantage to varimax rotation is that it better represents simple structure; a problem is that does not always clearly define factors. Most researchers tend to prefer varimax rotation to any other method available (Taylor, 1977, Rummel, 1970).

Quartimax rotation is useful in identifying and comparing a particular variable's variance and as a method of classifying variables by the factor to which they are most similar. It is useful in simplifying variable description.

An advantage to quartimax rotation is that it more precisely defines factors; a problem with quartimax rotation is that it tends to artificially increase the first factor's amount of explained variance (Rummel, 1970; Weiss, 1971).

Oblique rotations permit factors to be intercorrelated. Each factor is individually rotated to align with its representative cluster of variables. One advantage is that important variables which define factors can be identified by the factor loadings. Also, by fitting each factor separately, more precise cluster definition is allowed. Furthermore, the correlation between clusters can be interpreted to determine if the data are, in fact, orthogonal or oblique. In reality, most factors are correlated. One example is the education factor and the income factor. It is unrealistic to force these factors to be orthogonal or uncorrelated (Rummel, 1970; Weiss, 1971).

Oblique rotated factors are much more difficult to interpret than orthogonal rotated factors. Instead of only creating one matrix consisting of factor loadings, it is necessary to create two matrices; the primary structure matrix and the reference matrix. The primary structure matrix delineates pattern loadings. Pattern loadings measure the unique contribution each factor makes to each variable's variance. Because they measure variable dependence upon different factors, pattern loadings can be interpreted as regression weights. The reference matrix delineates structure loadings. Structure loadings represent the product moment correlations of the variables with the factors. The product moment correlations are used to rotate the intercorrelation matrix to oblique structure (Rummel, 1970).

One oblique rotation that is commonly used is promax rotation. Promax rotation is used in the United Way geo-demographic database system (O' Hare, 1986). First, it is used to rotate the factors to an orthogonal, varimax structure. This rotation is accomplished by taking the k^{th} power of each factor loading. The Procrustes technique is then used to calculate the least squares fit which adjusts the factors to represent oblique structure. The advantage to promax rotation is that factors will be more clearly identified. This is possible because variables with already high factor loadings will be artificially increased. Likewise, variables with low factor loadings will be decreased. A problem with this method is to determine what power k should be. The most commonly used value of k is 4.0 (Rummel, 1970).

The amount of explained variance that each variable contributes to each factor's formation is measured in the form of squared factor loadings. The table of standardized factor scores can be used as input into cluster analysis. Cluster analysis forms groups based on the amount of variance that each variable contributes towards factor formation.

The Use of Cluster Analysis

Cluster analysis is another commonly used data classification method (Punj and Stewart, 1983). Cluster analysis is a multivariate statistical procedure which divides a data set into homogeneous groups (Aldenderfer and Blashfield, 1984). The difference between cluster analysis and factor analysis is how an observation's variance is analyzed.

Factor analysis divides variance among several factors, while cluster analysis assigns total variance into discrete sets (Borger and Weiss, 1971). Cluster analysis is particularly useful for empirical studies because it does not make any prior assumptions about data structure except that it is orthogonal. Three common uses of cluster analysis are for market segmentation, identifying homogeneous buying groups, and to parsimoniously reduce the size of a data set (Punj and Stewart, 1983).

There are some potential problems that the user needs to be aware of when using cluster analysis. One of these problems is that cluster analysis is a heuristic, not an exact, method. It is not possible to subject clustering methods to statistical verification, as it is with factor analysis (Aldenderfer and Blashfield, 1984; Blashfield and Aldenderfer, 1978). Another problem is that there are many cluster analytic methods available. Each method produces its own unique solution. Thus, both inexperienced and experienced users tend to be confused about the properties of the various clustering methods. Further complicating this situation, there has not been any attempt to standardize the terminology used in naming clustering techniques (Aldenderfer and Blashfield, 1984; Punj and Stewart, 1983; Blashfield and Aldenderfer, 1978). As an example, there are six equivalent terms for single linkage cluster analysis (Blashfield and Aldenderfer, 1978).

There are two general types of cluster analysis methods available; iterative partitioning and hierarchical (Blashfield and Aldenderfer, 1978; Borger and Weiss, 1971). Iterative partitioning methods require the user to decide how many clusters are needed to classify the data set.

Based on this number, there are various iterative methods that are used to group the data into clusters. Hierarchical clustering methods forms clusters by successfully grouping the most similar entities together. The grouping process creates a tree like structure (Blashfield and Aldenderfer, 1978; Borger and Weiss, 1971).

Hierarchical clustering methods are the most commonly used. These methods have been extensively cited in the literature and their accuracy has been empirically verified (Lorr, 1983; Blashfield and Aldenderfer, 1978; Borger and Weiss, 1971). One advantage of hierarchical clustering methods is the imposition of the hierarchical or tree like structure. This structure makes cluster assignments more easily understood. With hierarchical agglomerative clustering, the observations are treated as independent entities at the lowest level of the hierarchy. As the hierarchical levels increase, the observations begin to merge into clusters. At the top of the hierarchy, all of the observations are merged into one large cluster. Hierarchical divisive methods begin with all observations being assigned to one large cluster and are divided into all observations are assigned to one member groups. (Aldenderfer and Blashfield, 1984; Lorr, 1983). Another advantage of using hierarchical clustering methods is that previous knowledge of the number of clusters needed to classify the data set is not needed, as it was with iterative partitioning methods.

Three problems with hierarchical clustering methods are in deciding when to stop the clustering procedure, the inability to reassign poorly grouped observations, and sensitivity of the clustering procedure to outliers in the dataset (Lorr, 1983).

Unfortunately, there is not any procedure or statistic that can be used to optimally determine how many clusters are needed to group a data set. There are two statistics that are helpful in deciding how many clusters are needed. The scree test, changes in cluster size and a subjective examination of the results to determine how many clusters are necessary to represent the variation in the data. Further, the structure of hierarchical clustering does not permit poorly grouped observations to be reassigned, and this deficiency cannot be corrected. Outliers, that are not within the normal range of observations, must be classified into clusters. Unfortunately, outliers may force seriously distorted cluster formation (Lorr, 1983; Borger and Weiss, 1971).

Hierarchical clustering methods can be classified into two types: divisive and agglomerative. The divisive methods begin with the data grouped into one large cluster and then subdivide into more clusters in a stepwise fashion. Even though there are a number of hierarchical divisive methods available, there have been very few examples of their application in the literature (Blashfield and Aldenderfer, 1978). For this reason, I will not be testing them. The hierarchical agglomerative methods begin with all the observations existing as isolated cases. Then, these isolated cases are grouped together in a stepwise fashion until all observations end up in one large cluster (Blashfield and Aldenderfer, 1978; Borger and Weiss, 1971). Because of their popularity among researchers, I will concentrate on using hierarchical agglomerative methods.

Of the approximately twelve hierarchical agglomerative clustering techniques available, four are commonly used.

They are single linkage, complete linkage, average linkage, and Ward's method (Aldenderfer and Blashfield, 1984; Punj and Stewart, 1983). Average linkage, single linkage and complete linkage are the most commonly used methods (Lorr, 1983). Ward's method is used in the United Way's geo-demographic database system and is also accepted among researchers (O'Hare, 1986; Borger and Weiss, 1971). Researchers have attempted to determine which of these clustering techniques is superior, but their results have not been conclusive (Aldenderfer and Blashfield, 1984; Lorr, 1983; Blashfield and Aldenderfer, 1978). This makes it all the more important to try different methods and compare the results, since it is not possible to prove one of these methods superior to the others in all cases.

Single linkage analysis requires that an observation be similar to at least one of the cluster members before it can be joined to the cluster. As more clusters are formed, the similarity criterion required for cluster membership is lowered, as the observations become less related (Punj and Stewart, 1983; Sokal and Sneath, 1963). This method is useful in detecting irregular and elongated clusters (Lorr, 1983). Further, it is the only method that is insensitive to monotonic transformations of the input data (Aldenderfer and Blashfield, 1984). One problem with single linkage cluster analysis is, however, that it often produces chained clusters. Chained clusters are long, straggly clusters that are poorly grouped (Aldenderfer and Blashfield, 1984; Lorr, 1983; Punj and Stewart, 1983).

An example of a chained cluster solution would be if the data set consisted of 100 observations was to be grouped into two clusters, one cluster containing 99 observations and the other only having one observation (Aldenderfer and Blashfield, 1984; Lorr, 1983; Punj and Stewart, 1983).

Complete linkage cluster analysis differs from single linkage cluster analysis because an observation must be similar to all of the members in the cluster, before it can be joined. Because of this requirement, the clusters are normally homogeneous in nature (Aldenderfer and Blashfield, 1984; Punj and Stewart, 1983).

Average linkage cluster analysis is viewed as a compromise between single linkage cluster analysis and complete linkage cluster analysis. This method calculates average similarity for the observation in question, and calculates an average for all the observations within the cluster. An observation is permitted to join to a cluster only if it meets the average level of similarity criteria. (Punj and Stewart, 1983; Aldenderfer and Blashfield, 1984). There are four different methods used to calculate average similarity. Of these methods, Sokal and Sneath's simple average linkage method is the most commonly used (Punj and Stewart, 1983). Sokal and Sneath's method is useful because it creates small, well separated clusters and is the least sensitive to outliers (Lorr, 1983).

Ward's method forms clusters by minimizing within-cluster variance. It differs from other clustering methods because it attempts to minimize the within groups sum of squares and the cross products matrix.

Ward's method tends to produced clusters of similiar size and shape (Aldenderfer and Blashfield, 1984; Punj and Stewart, 1983, Lorr, 1983). It is sometimes viewed as a special case of the average linkage method, because it uses the within-group sum of squares value to minimize distance within clusters (Punj and Stewart, 1983).

Iterative partitioning methods are not as commonly used as hierarchical methods. Most iterative partitioning methods use the following four step process (Anderberg, 1973).

1. Choose the number of clusters and partition the data set based on this number. Then, calculate each cluster's centroid.
2. Allocate each observation to the nearest cluster.
3. Recalculate cluster centroids after all observations have been assigned to a cluster.
4. Repeat step two and step three until no observations change cluster

(Anderberg, 1973).

One advantage to using iterative partitioning methods are their ability to reassign poorly formed clusters (Aldenderfer and Blashfield, 1984; Lorr, 1983). Another advantage is that the data matrix is computationally easier to calculate. Iterative partitioning can handle larger data sets than hierarchical methods (Aldenderfer and Blashfield, 1984; Lorr, 1983). A major problem with iterative partitioning methods is determining the optimal number of clusters with which to group the data. For a problem with 15 variables which will be grouped into three clusters, there are, for example, 217,945,728,000 possible cluster assignments that must be compared to determine the composition of the optimal number of clusters (Aldenderfer and Blashfield, 1984).

Obviously, from a computational standpoint, it would impossible to optimally solve most large problems. This is why the preceding four step heuristic process is necessary to try to estimate the correct number of clusters needed to classify the data set. This suboptimality results from the inability to empirically determine how many clusters should be created. Another problem with iterative partitioning methods is that they commonly result in suboptimal solutions. Normally, this happens because initial cluster positions are poorly selected (Aldenderfer and Blashfield, 1984).

The following chapter will be used to develop an intra-urban version of the United Way's geo-demographic database system. The two research hypotheses introduced in Chapter I will be tested. A list of the procedures used in hypothesis testing is included in Chapter I. You may need to use this chapter to provide an explanation of the characteristics of factor analysis and cluster analysis.

CHAPTER V

DATA ANALYSES

In this chapter, I discuss the step-by-step procedures necessary to develop a geo-demographic database system. I also test the two research hypotheses discussed in Chapter I. Hypothesis one involves examining the factoring and clustering techniques that are used to develop a geo-demographic database system. Specifically, the group assignments are hypothesized to be highly sensitive to the factoring and clustering methods employed. Test 1 and test 2 described in Chapter I were employed to examine the sensitivity of factor analytic and cluster analytic results to the particular methods utilized.

Hypothesis two involves examining the sensitivity of cluster results to the input variables used. My hypothesis is that if a data set used in an existing geo-demographic database system is merged with a data set developed using variables selected from previous factorial ecology and quality of life studies, there will be a significant improvement in the amount of explained variance. With an improvement in the amount of explained variance, cluster formation will be more precise and more easily determined. Test 3, described in Chapter 1, was employed to examine the sensitivity of cluster results to the input variables used.

Test 1. Comparing Commonly Used Factor Analytic Methods

Testing the sensitivity of factor analytic results to the factor analytic technique and rotation selected was the first test conducted. The following material describes the step-by-step procedures used in this sensitivity analysis.

Step 1.1 - Develop database for the Knoxville MSA using the United Way's geo-demographic database system variables.

In order to probe the sensitivity of the grouping to the type of procedures used, it was necessary to develop a database based on variables extracted from an existing geo-demographic database system. I was fortunate to be able to obtain information describing the variables and data classification techniques used to develop the United Way's geo-demographic database system. Even though their study was conducted at the inter-urban scale, nearly all of the variables and their classification techniques can be used as a basis of the present intra-urban study.

Using the United Way's variables (Appendix 1) and research methodology, I developed an intra-urban geo-demographic database for the Knoxville MSA at the census tract level. In 1980, the Knoxville MSA consisted of four counties: Anderson, Blount, Knox and Union. The total population of the MSA in 1980 was 476,517 (U.S. Bureau of the Census, 1983b). The database consisted 53 Census-derived variables for 111 census tracts.

These data would have been extremely time consuming to enter manually. I decided, instead, to extract the 53 variables directly from census tapes. ¹. For a discussion of the procedures necessary to access census tapes, see Appendix 3.

Step 1.2 - Replicating the United Way study at the intra-urban scale

The first analytical procedure after the data base was created was to replicate the United Way's factor analytic solution at the intra-urban level. The United Way's inter-urban geo-demographic data base system used a five factor Principal Components Analysis procedure with promax rotation. The first step was to use this same method to factor analyze the Knoxville data set created in step 1.1 (see Table 5.1). The total amount of explained variance that this solution provided was used as a benchmark against which to compare and evaluate other factoring methods and rotational schema in order to determine if the United Way's method best classified the data set or if it would be more appropriate to use another factoring and/or rotational method.

¹. Census tracts seven and thirty six were excluded from the study, because data were unavailable.

TABLE 5.1
 FACTOR ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS
 USING DEFAULT UNITED WAY SOLUTION OF
 PRINCIPAL COMPONENTS ANALYSIS
 PROMAX ROTATION ¹.

Factor	% of Explained Variance Ignoring Other Factors
1	17.947068
2	15.468461
3	9.412277
4	7.776669
5	4.880918
Total Explained Variance = 55.485393	

1. This was the same solution as was derived from use of 53 variables used to create the United Way's inter-urban geo-demographic database system for 111 of 113 census tracts.

Step 1.3 - Testing various factor analytic solutions

The next step in the comparative procedure was to vary the number of factors, factoring methods and rotations and then compare the factor analytic results. Both principal axis factor analysis and principal components analysis solutions were generated using three common rotational schemes -- varimax, quartimax, and promax. Each set of factor results was created using version 5.0 of the SAS Statistical Analysis System's software (SAS 1985). I allowed the SAS program to produce the solution which created the maximum number of factors possible without violating the Kaiser criterion. The Kaiser criterion requires each factor to explain at least one percent variance it is not considered to be significant. Then, using the amount of variance that each factor contributed as a guide, I varied the number of factors and examined the different solutions. To determine the number of factors that best classified the data set for each factoring technique and rotation, I examined the following criteria: 1) the scree plot; 2) the amount of explained variance contributed by each factor; 3) the change in the amount of variance that each factor contributed; and 4) and how the factor loadings greater than $|.4|$ delimited the dimensions of the data set. Then, I examined the results of each of these techniques to decide which method best classified the data set. Because of space limitations, the default solution and the best solution for each method are reported in Tables 5.2 - 5.7.

TABLE 5.2
FACTOR ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS
PRINCIPAL AXIS FACTOR ANALYSIS
VARIMAX ROTATION ^{1.}

Factor	% of Explained Variance
Default Solution	
1	16.170031
2	9.481205
3	5.361398
4	5.044893
5	4.685442
6	2.968841
7	1.992866
8	1.650065
9	1.087276
Total Explained Variance = 48.452018	
Best Solution ^{2.}	
1	14.601059
2	13.498864
3	6.978935
4	5.288817
5	2.881414
Total Explained Variance = 43.248868	

1. The study is based on using 53 variables previously used by the United Way of America in their inter-urban geo-demographic database system for 111 of 113 census tracts.

2. The best solution is determined by examining the scree plot, the amount of explained variance contributed by each factor, change in the amount of variance contributed by each factor, and how the factor loadings greater than |.4| identified the dimensions of the data set.

TABLE 5.3
FACTOR ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS
PRINCIPAL AXIS FACTOR ANALYSIS
QUARTIMAX ROTATION ^{1.}

Factor	% of Explained Variance
Default Solution	
1	20.116799
2	9.561581
3	8.462164
4	2.380548
5	2.150998
6	1.826526
7	1.556361
8	1.367898
9	1.029131
Total Explained Variance = 48.452018	
Best Solution ^{2.}	
1	18.003807
2	12.510514
3	7.712263
4	3.043376
Total Explained Variance = 41.269860	

1. The study is based on using 53 variables previously used by the United Way of America in their inter-urban geo-demographic database system for 111 of 113 census tracts.

2. The best solution is determined by examining the scree plot, the amount of explained variance contributed by each factor, change in the amount of variance contributed by each factor, and how the factor loadings greater than |.4| identified the dimensions of the data set.

TABLE 5.4
FACTOR ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS
PRINCIPAL AXIS FACTOR ANALYSIS
PROMAX ROTATION ^{1.}

Factor	% of Explained Variance
Default Solution	
1	19.384318
2	14.002048
3	10.536629
4	7.195370
5	8.293519
6	6.846165
7	5.991934
8	2.484444
9	2.938094
Total Explained Variance = 77.673522	
Best Solution ^{2.}	
1	17.154732
2	16.314530
3	9.459761
4	6.439676
5	5.675439
Total Explained Variance = 55.044138	

1. The study is based on using 53 variables previously used by the United Way of America in their inter-urban geo-demographic database system for 111 of 113 census tracts.

2. The best solution is determined by examining the scree plot, the amount of explained variance contributed by each factor, change in the amount of variance contributed by each factor, and how the factor loadings greater than |.4| identified the dimensions of the data set.

TABLE 5.5
 FACTOR ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS
 PRINCIPAL COMPONENTS ANALYSIS
 VARIMAX ROTATION ^{1.}

Factor	% of Explained Variance
Default Solution	
1	15.432052
2	12.037902
3	6.845497
4	4.920893
5	3.206980
6	2.801525
7	2.350238
8	1.427449
9	1.400810
Total Explained Variance = 50.223126	
Best Solution ^{2.}	
1	21.797763
2	9.750380
3	7.776746
4	2.953443
5	2.191907
6	1.861225
Total Explained Variance = 46.331464	

1. The study is based on using 53 variables previously used by the United Way of America in their inter-urban geo-demographic database system for 111 of 113 census tracts.

2. The best solution is determined by examining the scree plot, the amount of explained variance contributed by each factor, change in the amount of variance contributed by each factor, and how the factor loadings greater than .4 identified the dimensions of the data set.

TABLE 5.8
FACTOR ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS
PRINCIPAL COMPONENTS ANALYSIS
QUARTIMAX ROTATION ¹.

Factor	% of Explained Variance
Default Solution	
1	20.128871
2	9.867868
3	8.580009
4	2.670479
5	2.510698
6	2.008410
7	1.760910
8	1.558606
9	1.200274
Total Explained Variance = 50.223128	
Best Solution ².	
1	17.992356
2	12.447582
3	7.826704
4	4.011711
Total Explained Variance = 42.278832	

1. The study is based on using 53 variables previously used by the United Way of America in their inter-urban geo-demographic database system for 111 of 113 census tracts.

2. The best solution is determined by examining the scree plot, the amount of explained variance contributed by each factor, change in the amount of variance contributed by each factor, and how the factor loadings greater than .4 identified the dimensions of the data set.

TABLE 5.7
FACTOR ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS
PRINCIPAL COMPONENTS ANALYSIS
PROMAX ROTATION ^{1.}

Factor	% of Explained Variance Ignoring Other Factors
Default Solution	
1	18.234303
2	15.467032
3	9.043938
4	8.901353
5	5.439060
6	6.709278
7	6.786543
8	2.400616
9	4.313256
Total Explained Variance = 77.295379	
Best Solution ^{2.}	
1	17.947068
2	15.468481
3	9.412277
4	7.776669
5	4.880918
Total Explained Variance = 55.485393 ^{3.}	

1. The study is based on using 53 variables previously used by the United Way of America in their inter-urban geo-demographic database system for 111 of 113 census tracts.

2. The best solution is determined by examining the scree plot, the amount of explained variance contributed by each factor, change in the amount of variance contributed by each factor, and how the factor loadings greater than |.4| identified the dimensions of the data set.

3. This is the factoring method and rotation used in the United Way's inter-urban geo-demographic database system.

Step 1.4 A Comparison of the factor analytic results

After examining the results from the factor analyses, there were two results that were quite similiar. The United Way's five factor Principal Components Analysis promax rotation and the five factor Principal Axis Factor Analysis promax rotation were the solutions that provided the highest amount of explained variance without violating the constraints discussed in Chapter 4. To determine if the solutions had a similiar structure, the coefficient of congruence for each of the five factors was calculated. The purpose of calculating the coefficient of congruence for each of the factors was to compare the factor loadings obtained by the two different methods (Johnston, 1980)

In order to calculate the coefficient of congruence, I used Johnston's (1980) formula.

$$G_{ij} = \frac{L_{ki} L_{kj}}{\text{SQR}(L^2_{ki} L^2_{kj})}$$

L_{ki} represents the factor loadings for variable K for matrix i.

L_{kj} represents the factor loadings for variable K for matrix j.

G_{ij} is the coefficient of congruence between the two matrices.

The coefficient of congruence statistics for all five factors is reported in Table 5.8.

TABLE 5.8
COEFFICIENT OF CONGRUENCE STATISTICS FOR KNOXVILLE MSA CENSUS TRACTS
COMPARING
FIVE FACTOR PRINCIPAL AXIS FACTOR ANALYSIS PROMAX ROTATION SOLUTION
AND
PRINCIPAL COMPONENTS ANALYSIS PROMAX ROTATION SOLUTION

FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
.9123	.8772	.9474	.8772	.7718

All five coefficient of congruence statistics are extremely high. They should be interpreted as correlations. These statistics demonstrate that the factoring method and rotation does not seem to make a significant difference in the amount of explained variance or in factor structure. If it did, these statistics would display low values, indicating that there was a significant difference between the factor loadings.

Based on my examination of the factor analytic results, and the coefficients of congruence, I decided that the best method for analyzing this data set was the same as used by the United Way in an inter-urban context -- a five factor Principal Components Analysis using promax rotation. This solution maximizes explained variance without violating the factor analytic constraints discussed in detail in Chapter IV.

It was clear, however, that there was a significant improvement in the amount of explained variance if promax rotation was used in comparison to varimax and quartimax rotations.

Because of the improvement in overall explained variance, I concluded that the technique used in factor rotation is important in determining the results.

The factor loadings from the Principal Components solution using promax rotation are reported and described in Table 5.9. Factor loadings less than |0.4| are excluded because they are not considered to be significant. The standardized factor scores generated for each census tract were used cluster analysis input where the sensitivity of group assignments to the clustering techniques employed could be tested.

Discussion of Table 5.9 Factor Loadings Within the Knoxville MSA

Factor one indicates a "white, blue collar, middle class family" dimension of the data set. The percentage of white population has a extremely high factor loading of .86. Five of the six variables measuring percentage of population at different age levels have high factor loadings. The percentage of population less than five, percentage of population between 10 and 19, and the percentage of population between 35 and 54 have factor loadings $\geq .72$ offering further evidence of the familial aspects of this factor. Three of the five income variables have high factor loadings. These loadings correspond with income levels that would be expected of middle class households. Blue collar occupational variables have high factor loadings and the high loading on percentage of population greater than 25 completing less than four years of high school provides additional proof of the blue collar nature of this group.

TABLE 5.9
FACTOR LOADINGS OF KNOXVILLE MSA CENSUS TRACTS
PRINCIPAL COMPONENTS ANALYSIS PROMAX ROTATION
BASED ON STANDARDIZED REGRESSION COEFFICIENTS

VARIABLES	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
% OF OCCUPIED HOUSING UNITS	.5503	.4310		.4321	
% OF URBAN POPULATION					.6378
% OF BLACK POPULATION					.8107
% OF WHITE POPULATION	.8599				
% OF INDIAN POPULATION					
% OF ASIAN POPULATION			.6204		
RATIO OF NWKRS TO WORKERS			-.6566		
% EMP. OUTSIDE COUNTY OR ST.	.7053		.4152		
% FEMALE HEADED HH WITH KIDS					.8119
% OF NONFAMILY HOUSEHOLDS		.4797	.4687	.6612	
% OF POP. GREATER THAN 65				.7910	
% OF POP. LESS THAN 5	.8202				
% OF POP. BETWEEN 10 AND 19	.7242		.5220		
% OF POP. BETWEEN 20 AND 34	.5153		.7637		
% OF POP. BETWEEN 35 AND 54	.8488				
% OF POP. BETWEEN 55 AND 64	.5256			.4759	
% OF POP. 25+ COMPLET. < 4YRS HS	.6870	.4934		.4844	
% OF POP. 25+ COMPLET. 1 & 3 YRS COLL.		.8647			
% OF POP. 25+ COMPLET. 4+ YRS COLL		.9181			
% OF POP. WITH INC. < THAN \$5000				.7006	
% OF POP. WITH INC. \$7,500 & \$9,999				.7308	
% OF POP. WITH INC. \$10,000 & \$14,999	.4496			.6313	
% OF POP. WITH INC. \$15,000 & \$24,999	.7670				
% OF POP. WITH INC. \$25,000 & \$49,999	.5253	.8641			
% OF POP. WITH INC. > THAN \$50,000		.8481			
% OF IND. WITH INC. < THAN \$5000			.8622		
% OF IND. WITH INC. \$7,500 & \$9,999			.7774		
% OF IND. WITH INC. \$10,000 & \$14,999		.7462		.4906	
% OF IND. WITH INC. \$15,000 & \$24,999		.8472			
% OF IND. WITH INC. \$25,000 & \$49,999		.8798			
% OF IND. WITH INC. > THAN \$50,000		.5779			
% EMPLOYED AS PRIVATE WAGE EARNERS	.5203				
% EMPLOYED IN STATE GOVERNMENT			.8009		
% EMPLOYED IN LOCAL GOVERNMENT	.6578				
% SELF EMPLOYED	.6230	.4189			
% EMP. IN AG., FORESTRY AND MINING	.6085				-.4807
% EMP. IN CONSTRUCTION	.8871				
% EMP. IN MANUFACTURING	.9270				
% EMP. IN TRANS-PUBLIC UTILITIES	.7120				

TABLE 5.9 (Continued) 1.

VARIABLES	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
% EMP. IN RETAIL TRADE	.5655				
% EMP. IN F.I.R.E.		.7758			
% EMP. IN SERVICES		.8152			
% EMP. IN PUBLIC ADMINISTRATION	.4537	.5382			
% EMP. IN MANAGEMENT		.8526			
% EMP. IN TECH/SALES	.5276	.5987			
% EMP. IN SERVICE OCC.			.5484		
% EMP. IN PRECISION PROD.	.8711				
% EMP. IN OPERATOR-LABORER POSITIONS	.8772	-.4976			
% OF OCCP. RENTAL HOUSING UNITS		.4040		.6564	
% OF POP. 5+ IN SAME HOME AS IN 1975	.8405				
% OF POP. 5+ MOVE WITHIN STATE	.6186				
% OF POP. 5+ MOVED FROM OTHER STATE		.4089	.4589		
% OF POP. 5+ ENTERED COUNTRY			.7727		
% OF POP. BELOW 75% OF POVERTY LEVEL		-.5175			
% OF POP. BETWEEN 75% & 125% OF POV.	.4255	-.4035		.6020	
% OF POP. BETWEEN 125% & 150% OF POV.	.4493			.5400	
% OF POP. AT OR ABOVE 200% OF POV.	.6706	-.5600			
% OF POP. OF SPANISH ORIGIN				.4136	.4758
VARIANCE EXPLAINED BY EACH FACTOR IGNORING OTHER FACTORS					
FACTOR 1	17.947068				
FACTOR 2	15.488481				
FACTOR 3	9.412277				
FACTOR 4	7.766869				
FACTOR 5	4.860918				
TOTAL VARIANCE	55.472573				

1. The study is based on using 53 variables previously used by the United Way of America in their inter-urban geo-demographic database system for 111 of 113 census tracts.

Factor two indicates a "white collar singles and small family" dimension of the data set. The variable measuring the percentage of non-family households has a significant factor loading. All three of the educational attainment variables have significant factor loadings. The loadings measuring the percentage of population over 25 completing one to three years of college education and the percentage of population over 25 completing four or more years of college education are $\geq .86$ demonstrating the importance of higher education for this factor. The factor loadings on variables measuring family income indicate that this second group has moderate to high income. Four of the five variables measuring individual income have significant factor loadings indicating that this factor measures non-family households. The factor loadings measuring white collar employment are significant, which supports the results obtained using the educational attainment variables, as well as the income variables.

Factor three indicates a "university" dimension of the data set. There are significant loadings on the variables that measure the percentage population of Asian origin, percentage of population of Spanish origin, and the percentage of the population entering the population entering the United States from other nations. These ethnicity variables may reflect people entering the United States to study at the University. Variables measuring the percentage of population between 10 and 19 and the percentage of population between 20 and 34 have high factor loadings indicating the existence of teenage and young adult population, which would be expected in an university environment.

The interpretation of factor three as a university dimension is further confirmed by examining employment variables. Employment variables with high factor loadings are the percentage of population employed in state government and percentage employed in service occupations. In the Knoxville MSA, the University of Tennessee, a state supported institution, employs many undergraduate and graduate students. Further, many students supplement their income with part time service positions, which is reflected in this factor loading.

Factor four indicates a "elderly" dimension of the data set. There is a high factor loading for the percentage of population living in non-family households. The loadings of the income variables indicates that income levels which are highly correlated with the fourth factor. The factor loadings for variables measuring poverty further confirm the low income of this group.

Factor five indicates a "black, single parent urban" dimension of the data set. There is a high factor loading for the percentage of black population. The high factor loading on urban population indicate that the distribution of black population is related to levels of urbanity. The high factor loading on female headed households with children is a unfortunate characteristic of many black, urban families.

Test 2. Comparing Commonly Used Clustering Methods

The purpose of the second test is to examine the sensitivity of group assignments to the type of clustering techniques employed.

This test uses as input the standardized factor scores created as a result of the factor analytic procedures. In order to compare the effect of using different clustering techniques, I used four commonly used clustering methods -- Average, Complete, Single, and Ward's -- to assign the census tracts to groups. Ward's method was used to group the standardized factor scores generated by the United Way's inter-urban geo-demographic database system. I was also interested in the performance of Ward's method, vis-a-vis the other clustering methods in an intra-urban test of the United Way's data set.

Step 2.1 Comparing four different clustering methods applied to the five factor Principal Components Analysis promax rotation solution

Using the standardized factor scores generated from the five factor Principal Components Analysis promax rotation solution, I used the four clustering methods to group the data. For an extensive discussion of the properties of these clustering methods, see Chapter IV. To examine changes in the number of clusters and cluster composition, I relied on the following criteria: 1) the scree plot; 2) changes in taxonomic distance; 3) my ability to interpret the results in light of my knowledge of the Knoxville MSA; and 4) mapping the census tract assignments. I compared each method's ability to assign census tracts into meaningful clusters. The results are reported and compared in Table 5.10 (page 78).

Step 2.2 - Comparing the cluster analysis results

After analyzing and comparing the number of cluster's created, tracts assigned to particular groups, and the maps produced from the application of each of these four methods, I feel that Ward's method best grouped the United Way's data set. The results were the easiest to interpret and seemed to make intuitive sense given my knowledge of the study area. With Ward's method, all census tracts are assigned to a group. Outliers were created using these other clustering routines. From the marketer's perspective, it is desirable to assign every population group to a cluster, even though that group might not be similar to the others assigned to the same cluster. However, from the academic perspective, outliers represent divergent population groups. Because of their uniqueness, it may do violence to overall understanding to assign such outliers to existing groups. Such is the tension between the academic and marketing purposes to which geo-demographic systems might be put. For a map of census tract assignments using Ward's clustering method, see Figures 5.1 - 5.4 (pages 79-86).

The last step in the grouping process is cluster identification. In order to identify each cluster, I examined the standardized factor score means', standard deviations, and their distribution for each cluster see Table 5.11 (page 87). A negative mean score means that the population group has a lower than average representation within a particular cluster and vice versa for a positive mean score.

TABLE 5.10
CLUSTER ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS USING
PRINCIPAL COMPONENTS ANALYSIS PROMAX ROTATION
BASED ON STANDARDIZED FACTOR SCORES 1. 2.

CLUSTER	NUMBER OF MEMBERS	CLUSTER	NUMBER OF MEMBERS
AVERAGE LINKAGE CLUSTERING METHOD		COMPLETE LINKAGE CLUSTERING METHOD	
1	18	1	10
2	12	2	9
3	2	3	14
4	24	4	2
5	2	5	2
6	9	6	10
7	13	7	13
8	2	8	4
9	2	9	14
10	4	10	4
11	6	11	5
12	2	12	10
13	3	13	3
14	3	14	2
15	2	15	2
OUTLIERS	7	16	3
SINGLE LINKAGE CLUSTERING METHOD		OUTLIERS	4
		WARD'S CLUSTERING METHOD	
1	1	1	18
2	2	2	5
3	11	3	14
4	2	4	10
5	7	5	13
6	15	6	13
7	3	7	13
8	6	8	11
9	16	9	14
10	4		
11	10		
12	14		
13	2		
OUTLIERS	17		

1. Hierarchical clustering requires all observations to be assigned to a cluster. SAS output is structured to let the interpreter decide how many clusters should be formed. In situations where the change in taxonomic distance was extreme, some census tracts were not assigned if it would have necessitated creating one member clusters.
2. The results are based on standardized factor scores created by factor analyzing the United Way's data set for 111 of 113 census tracts.

FIGURE 5.1
CLUSTER ASSIGNMENTS FOR ANDERSON COUNTY CENSUS TRACTS
USING UNITED WAY DATA SET

CLUSTER ASSIGNMENTS FOR ANDERSON COUNTY CENSUS TRACTS USING UNITED WAY DATA SET

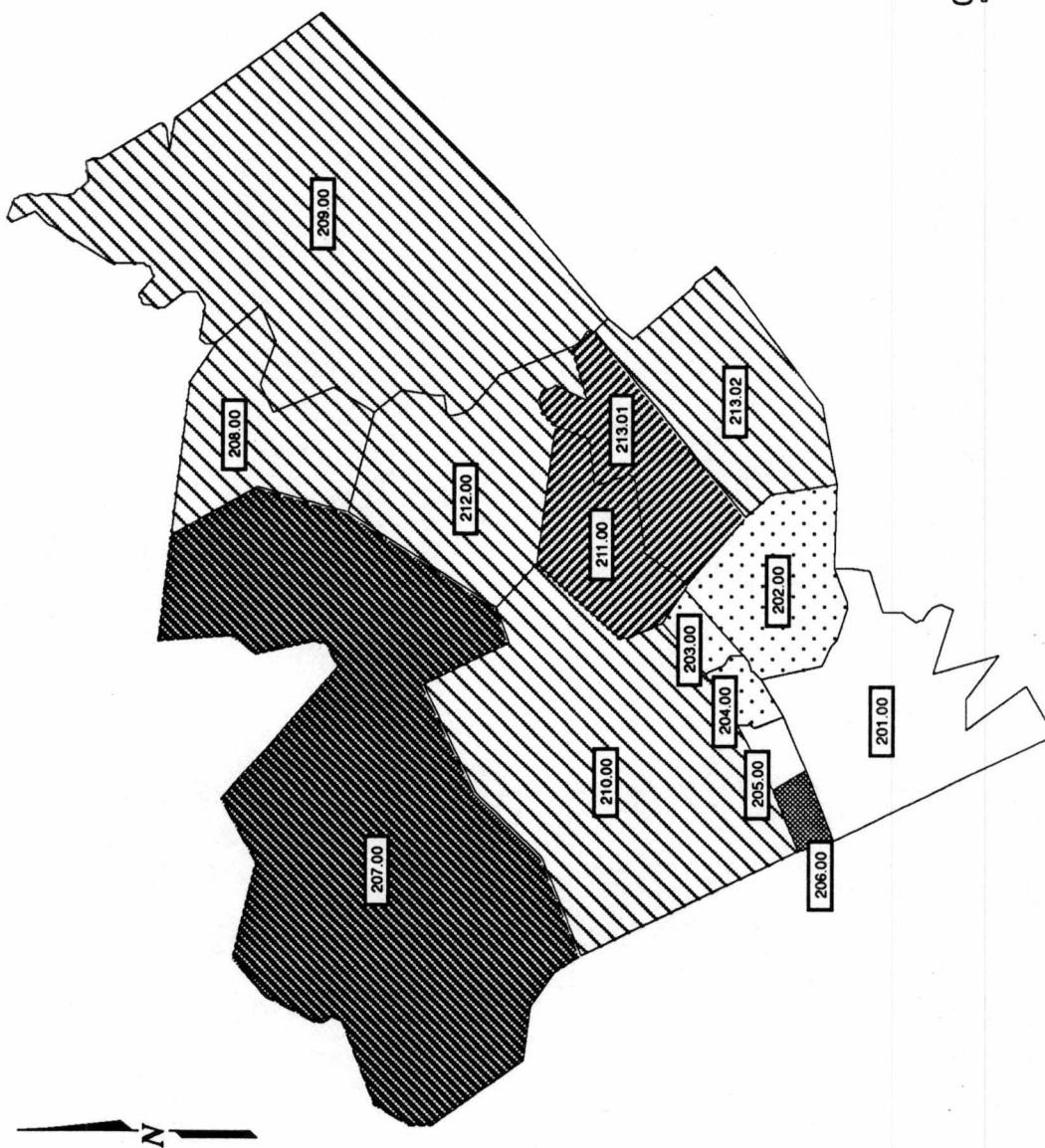
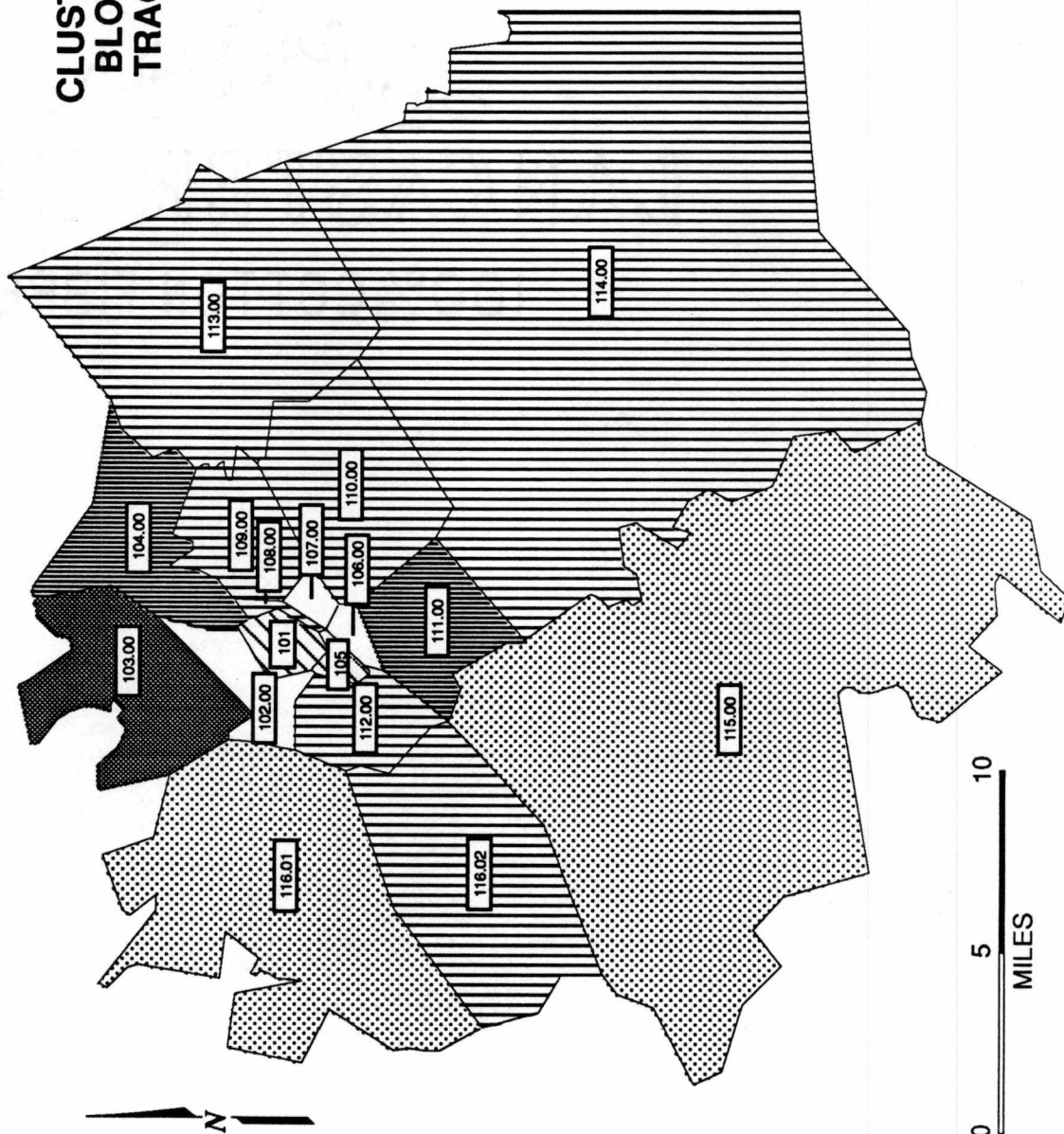


FIGURE 5.2
CLUSTER ASSIGNMENTS FOR BLOUNT COUNTY CENSUS TRACTS
USING UNITED WAY DATA SET

CLUSTER ASSIGNMENTS FOR BLOUNT COUNTY CENSUS TRACTS USING EXTENDED DATA SET



- ASSIGNMENTS**
- CLUSTER 1
 - CLUSTER 2
 - CLUSTER 3
 - CLUSTER 4
 - CLUSTER 5
 - CLUSTER 6
 - CLUSTER 7
 - CLUSTER 8
 - CLUSTER 9
 - CLUSTER 10
 - CLUSTER 11

FIGURE 5.3
CLUSTER ASSIGNMENTS FOR KNOX COUNTY CENSUS TRACTS
USING UNITED WAY DATA SET

CLUSTER ASSIGNMENTS FOR KNOX COUNTY CENSUS TRACTS USING UNITED WAY DATA SET

ASSIGNMENTS

- CLUSTER 1
- CLUSTER 2
- CLUSTER 3
- CLUSTER 4
- CLUSTER 5
- CLUSTER 6
- CLUSTER 7
- CLUSTER 8
- CLUSTER 9
- NO DATA

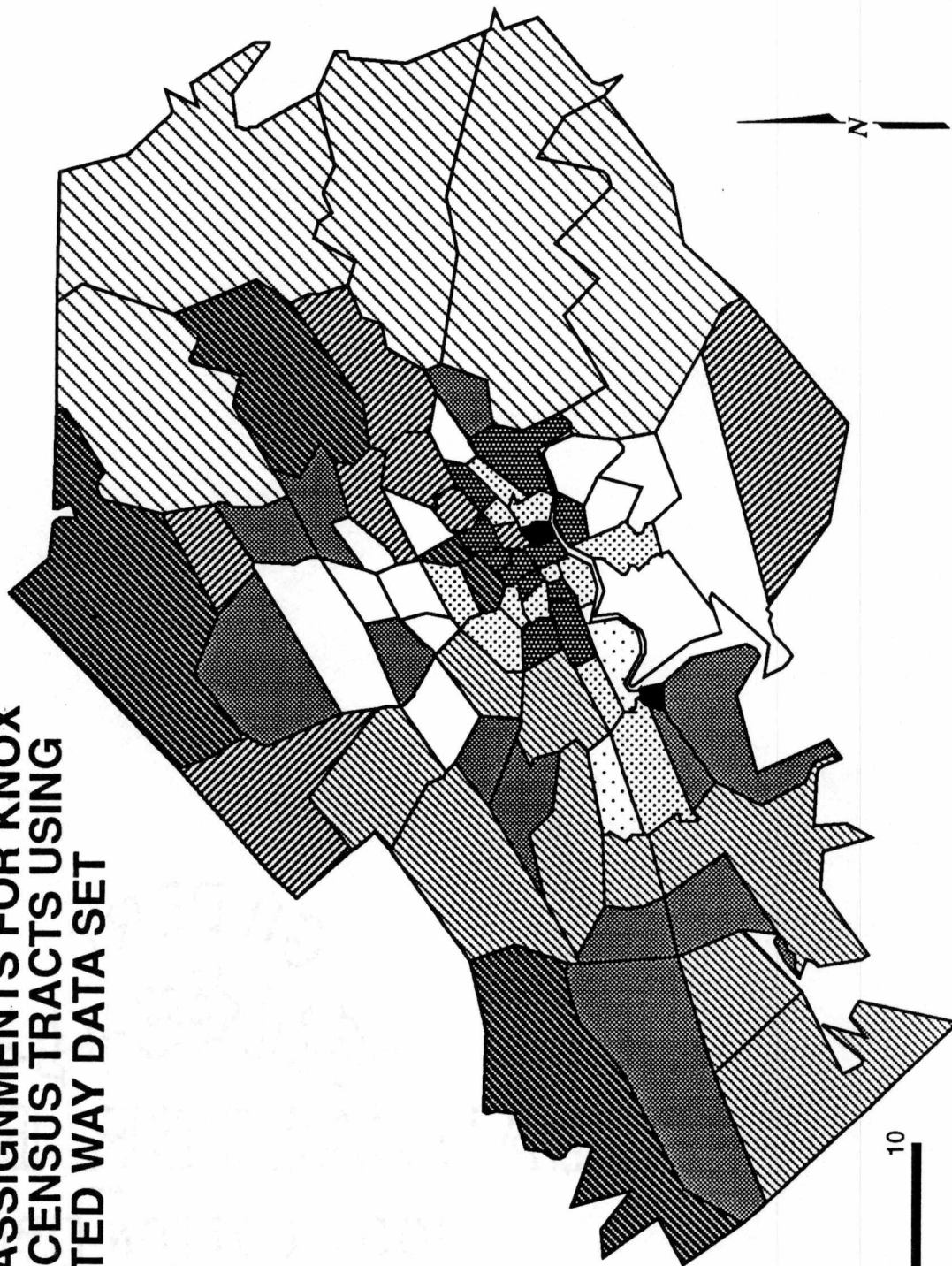
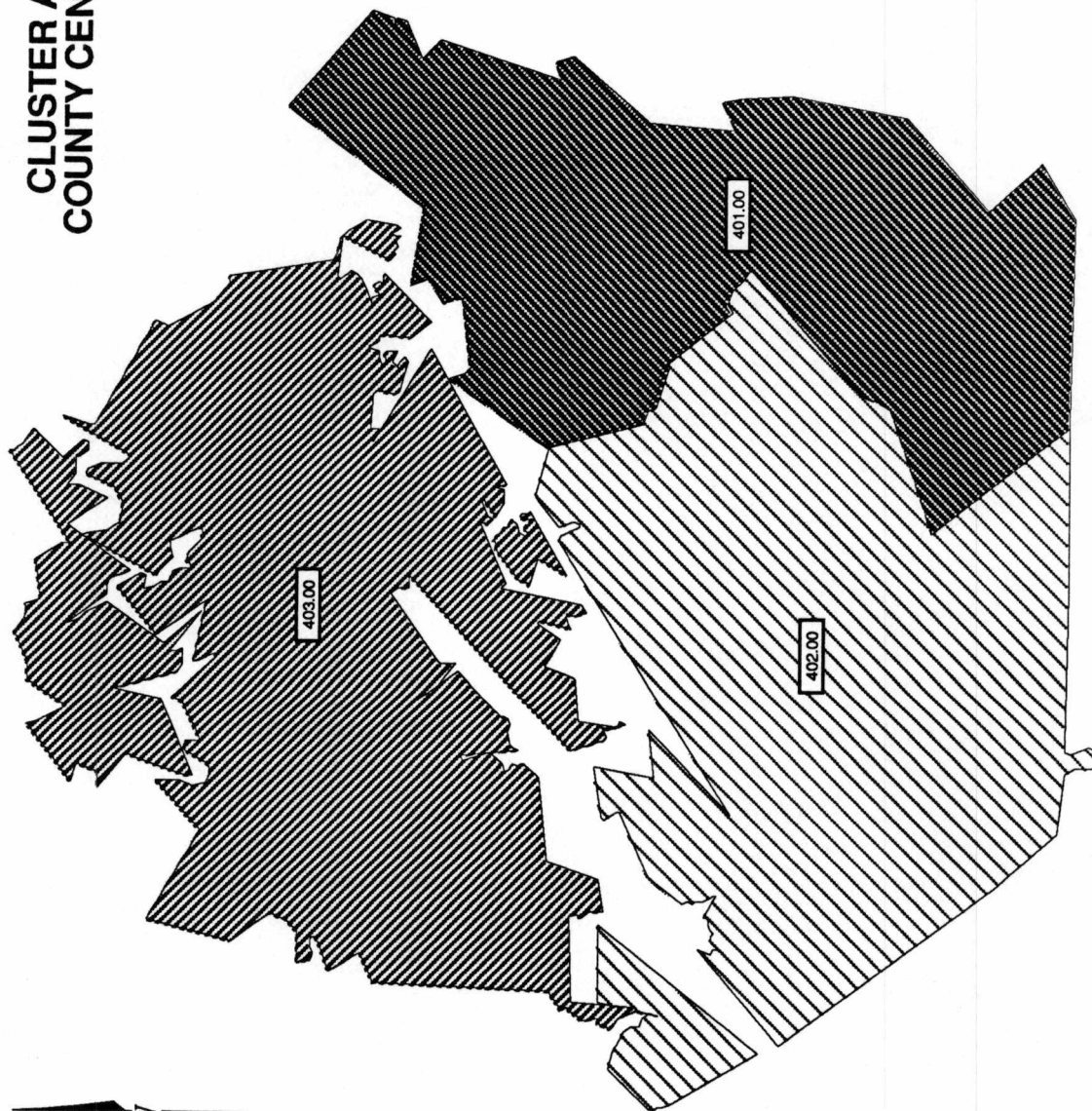


FIGURE 5.4
CLUSTER ASSIGNMENTS FOR UNION COUNTY CENSUS TRACTS
USING UNITED WAY DATA SET

CLUSTER ASSIGNMENTS FOR UNION COUNTY CENSUS TRACTS USING UNITED WAY DATA SET



ASSIGNMENTS

-  CLUSTER 3
-  CLUSTER 7
-  CLUSTER 8

0 5
MILES



TABLE 5.11
UNITED WAY'S DATA SET CLUSTER MEAN'S AND STANDARD DEVIATIONS
FOR KNOXVILLE MSA CENSUS TRACTS ¹.

Cluster	Factor1	Factor2	Factor3	Factor4	Factor5
Cluster Mean's					
1	0.0159	0.2794	-0.0161	0.1442	0.2260
2	-0.2241	1.9037	-0.0059	0.3195	-0.5906
3	1.4683	-0.4200	-0.1452	0.9460	-0.2996
4	1.4522	1.6585	0.8119	-0.6407	0.8200
5	-0.3778	-0.3700	1.1892	1.0405	1.8013
6	-0.2269	0.1917	-0.3645	-1.2089	-0.4470
7	-0.8539	-0.5797	-0.7523	-0.8570	-0.8432
8	0.0976	-0.3474	-0.4985	0.3221	-0.8538
9	-1.0522	-0.7355	-0.0803	0.0735	-0.1087
MAXIMUM	1.4683	1.9037	5.7744	2.2353	3.0368
MINIMUM	-0.8539	-0.7355	-0.7783	-1.5204	-0.9409
Cluster Standard Deviations					
1	0.5027	0.2264	0.3029	0.3083	0.4139
2	0.5022	0.5392	0.4203	0.6642	0.4310
3	0.4342	0.3749	0.2800	0.8905	0.3048
4	0.5191	0.8624	0.6390	0.8490	0.5835
5	0.6635	1.3809	2.1463	1.1433	1.2661
6	0.3733	0.5649	0.2843	0.5375	0.4328
7	0.4502	0.2987	0.1475	0.3446	0.2770
8	0.4756	0.8892	0.1875	0.6446	0.2894
9	0.5417	0.1769	0.5223	0.3373	0.4477

1. The source for this table was the standardized factor scores produced using the 57 variables extracted from the United Way's inter-urban geo-demographic database system.

CLUSTER 1

An examination of the mean factor scores for cluster one reveals an older, lower middle class white collar population group with elderly present. The positive score for factor two, the white collar singles and family dimension, factor four, the elderly dimension, and factor five, the black, single parent urban dimension enable me to make this determination. The distribution of tracts assigned to cluster one identify post WWII suburban tract housing areas in south Knoxville, north Knoxville, southwest Anderson County and central Blount county. Some examples of cluster one are tracts 22,23,34,35 and 56.01 in south Knoxville, and south Knox County.

CLUSTER 2

The mean factor scores for cluster two indicate that this is an older, upper middle class and wealthy population group. There is an extremely high mean score for factor two, the white collar singles and family dimension, and for factor four, the elderly dimension. Factor three, the university dimension's mean score near zero may indicate children of cluster residents attending a college or university and living at home. The other two factors have negative scores indicating that Cluster two's population is not identified by these factors. Some examples of cluster two are tract 25 - Sequoyah Hills and tract 45 -West Hills in the City of Knoxville.

CLUSTER 3

The mean factor scores for cluster three seem to indicate a predominantly blue collar population with some elderly present.

Both factor one, the blue collar family dimension, and factor four, the elderly dimension, have positive mean factor scores. Considering the distribution of tracts assigned to this cluster, it seems to delimit the mining and farming population group located in east Knox County, southwest Union County, western Blount County, and eastern Anderson County. One example of cluster three is tract 65, the Mascot mining community of east Knox County.

CLUSTER 4

With significant factor scores for four of five factors, this cluster seems to identify a younger, mixed population group. Blue and white collar families, singles, university, and urban blacks would be expected to exist in the group as indicated by the positive mean factor scores. The distribution of census tracts assigned to this cluster identifies a rural/suburban fringe area located in west Knox County, and northern Blount County. This is one area of new residential growth. Two examples of cluster four are tracts 58.01 and 58.02; the Concord community of west Knox County.

CLUSTER 5

With significant mean factor scores for factor three, the university dimension, factor four, the elderly dimension, and for factor five, the black, single parent urban dimension, it appears that this is the urban poverty group. University students, elderly, and black, single parent households in urban areas would be expected. University students are grouped with other impoverished groups because they have similar income characteristics. The distribution of census tracts assigned to cluster five indicate urban slum areas within the MSA.

Public housing projects exist within these tracts as well as the oldest and poorest quality housing. Some examples of cluster five are the Vestal community of south Knoxville (tract 24), and the University of Tennessee and Ft. Sanders Area (tracts 9 and 11).

CLUSTER 6

Factor two, the white collar singles and family dimension, was the only factor that had a positive score. After examining the distribution of census tracts assigned to cluster six, it seems to represent a white collar middle to upper middle class suburban fringe neighborhoods. This cluster is located in east Knoxville, north Knoxville and the Fountain City area, west Knoxville, west Knox County, and in southwest Anderson County. Some examples of this cluster are tracts 46.04 and 57.03 in the Farragut community of west Knox County.

CLUSTER 7

For cluster seven, none of the five factors have a positive mean score. This indicates that these factors do not reflect the socio-economic characteristics of this group. The distribution of census tracts assigned to this cluster delimits older middle class neighborhoods that have maintained their value. They are located in northeast Knoxville, southern Knox County, northern Union County, eastern Anderson County, and central Blount County. Two examples of cluster seven are tracts 211 and 213.01 representing the Lake City community in Anderson County.

CLUSTER 8

The mean factor scores for cluster eight seem to indicate that it has characteristics similiar to cluster four.

However, the mean score for factor four, the elderly dimension, is larger than the mean score for factor one, the white, blue collar middle class family dimension indicating that the elderly are more prevalent than blue collar families. Cluster eight seems to represent a rural, sparsely settled elderly and lower middle class population group existing in northern Knox County, eastern Union County, eastern Blount County, and northwestern Blount County. One example of cluster eight is tract 401, in Union County.

CLUSTER 9

Factor four, the elderly dimension, is the only factor with a positive mean score. The distribution of tracts assigned to cluster four indicates an older, lower middle class town population. Cluster nine is found in and around the central business district of Knoxville, and in central Blount County within the twin towns of Maryville/Alcoa. In such tracts is a lower middle class population which lives in older homes. There would be some mixed industrial and commercial activities within these tracts.

The description of each cluster provides only a brief understanding of cluster composition. To delve much more deeply into the nature of these clusters at this point would be premature because I have not yet decided if United Way's data set is the best one available to cluster the population. It is hypothesized that an extended data set employing variables derived from previous factorial ecology and quality-of-life studies in addition to those used by the United Way would do a better job of grouping the tracts into identifiable clusters.

Discussion of Hypothesis 1

Hypothesis 1 was tested to determine the sensitivity of group assignment to the factoring and clustering techniques employed. The results indicate that I should accept the hypothesis that the methods chosen, especially the rotational schema selected, are extremely important in determining what the results will be. Even though I accepted the hypothesis, the inherent subjectivity and context specificity of both factor and cluster analysis dictate that the procedures used above should be followed every time a researcher uses these methods. That is, there is no clearly superior set of classificatory procedures which should be used in all cases.

The next step in the analysis is to test the sensitivity of factor and cluster results to changes in the data set employed. This will be examined in the following section.

Test 3. Sensitivity of Clusters to the Input Variables Used

After determining the sensitivity of group assignment to the factor and clustering techniques employed, I explored the importance of the specific variables used in the creation of geo-demographic systems. My hypothesis is that a system will better classify areal units if the variables used to create a "modern" geo-demographic database are merged with variables used in classical urban ecological and quality-of life studies discussed in Chapter III.

To test this hypothesis, I created a new data set by merging the variables employed in the United Way system (Appendix 1) with an additional data set argued to be of importance in previous factorial ecology and quality-of-life studies (Appendix 2).

Based on this merged and extended data set, I developed an extended geo-demographic database system. I hypothesized that this extended data set should increase the amount of explained variance, and develop more precise cluster groupings than the United Way's data set. I subjected this extended data set to a factor analysis using the procedure and rotational scheme that worked best with the variables employed in the United Way's system. Then I compared the results of the extended system, to the results obtained from the application of the United Way system.

The next step subjected the standardized factor scores to cluster analysis. I used the clustering method (Ward's) that best grouped the United Way data set. I compared cluster groupings of the merged data set to those produced using the United Way data set.

Step 3.1 Developing the Extended Data Set

After examining many factorial ecology and quality of life studies, I felt that the United Way's data set could be improved. Twenty three variables determined to be of importance in previous factorial ecology/quality-of-life studies were added to the previously collected United Way data set. For a listing of the additional variables see Appendix II.

To create the extended data set, I used the 57 United Way variables and added 23 additional variables. This extended data set included two population variables, twelve housing variables, six employment variables, two income variables, and one mobility variable. I wished to determine if the amount of explained variance could be significantly improved and if the dimensions of the data set would be more easily defined if these variables were added. It is important to examine the change in amount of explained variance, because this statistic will indicate the amount of improved explanation that these additional variables provide. If there is only a slight increase in explained variance (say, less than 10%) it would probably not warrant using the extended system. However, if there is a significant improvement it will warrant using the extended system.

Step 3.2 - Factor Analyze the Extended Data Set

Using Principal Components Analysis with promax rotation, which was previously determined to best classify the United Way's data set, I conducted factor analysis on the extended data set. I varied the number of factors to be extracted from the data set between five and eight. To determine the number of factors that best classified the data set for this factoring technique and rotation, I examined the scree plot, the amount of explained variance contributed by each factor, the change in the amount of variance that each factor contributed, and how the factor loadings greater than $|.4|$ delimited the dimensions of the data set.

Then, I examined the results of each of these techniques to decide which method best classified the data set. The five factor solution seemed to be the best.

Table 5.12 lists and compares amount of variance that can be explained by five factors produced using the United Way's data set and the five factor solution produced with the extended data set. Table 5.13 lists the amount of variance that is contributed by each factor. It is necessary to generate Table 5.12 to measure the amount of variance that the 23 variables contribute independently because there is not an empirical test that can be applied to measure improvement in the performance of two different data set's applied to the same study area for classificatory purposes (Harman 1987). Table 5.14 lists the factor loadings that result from factor analyzing the 23 variables that are added to the United Way's data set alone. The factor loadings resulting from factor analyzing the extended data set are listed in Table 5.15 (pages 100-102). The standardized factor scores from Table 5.15 were used as input into cluster analysis which is discussed in step 3.3 (page 108).

An examination of Table 5.11 shows that there was a dramatic improvement in the amount of explained variance with the extended data set as compared to the United Way's data set. The amount of explained variance contributed by factor five nearly tripled. The amount of explained variance contributed by the other four factors increased approximately four percent. Only 55% of the variation in the United Way's data set was explained with the five factor solution.

TABLE 5.12
COMPARISON OF FACTOR ANALYTIC SOLUTIONS
FOR KNOXVILLE MSA CENSUS TRACTS
BETWEEN THE UNITED WAY'S DATA SET AND THE EXTENDED DATA SET
USING PRINCIPAL COMPONENTS ANALYSIS PROMAX ROTATION

Factor	% of Explained Variance Ignoring Other Factors	Factor	% of Explained Variance Ignoring Other Factors
United Way's Data set Best Solution ¹ .		Extended Data Set Best Solution ¹ .	
1	17.947068	1	22.465217
2	15.468481	2	20.968876
3	9.412277	3	13.136628
4	7.776869	4	11.731152
5	4.880918	5	11.243705
Total Explained Variance = 55.485393		Total Explained Variance = 79.545578	

1. The best solution is determined by examining the scree plot, the amount of explained variance contributed by each factor, change in the amount of variance contributed by each factor, and how the factor loadings greater than |.4| identified the dimensions of the data set.

TABLE 5.13
 FACTOR ANALYSIS OF KNOXVILLE MSA CENSUS TRACTS
 BASED UPON 23 VARIABLES EXTRACTED FROM PREVIOUS
 FACTORIAL ECOLOGY AND QUALITY-OF-LIFE STUDIES
 USING PRINCIPAL COMPONENTS ANALYSIS PROMAX ROTATION ¹.

Factor	% of Explained Variance Ignoring Other Factors
1	7.193135
2	4.682429
3	3.506116
4	3.549648
5	2.509314
Total Explained Variance = 21.440642	

1. These 23 variables chosen from previous factorial ecology and quality-of-life studies are tested separately to determine the amount of variance that is contributed to each factor. These 23 variables will be merged with the 57 variables used in the United Way data set to create the extended data set.

TABLE 5.14
FACTOR LOADINGS FOR KNOXVILLE MSA CENSUS TRACTS
BASED UPON 23 VARIABLES EXTRACTED FROM PREVIOUS
FACTORIAL ECOLOGY AND QUALITY-OF-LIFE STUDIES
BASED ON STANDARDIZED REGRESSION COEFFICIENTS ¹.

VARIABLES	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
% OF MARRIED COUPLE FAMILIES	.8711				
% OF POPULATION < 15 OR > THAN 85	.8543				.5074
% OF UNEMPLOYED INCLUDED IN WORKFORCE			.9046		
% OF EMPLOYED POP. WORKING PART TIME		.4114	.9310		
% OF HH. ON PUBLIC ASSISTANCE				.6150	.7378
% OF HH. ON SOCIAL SECURITY	.8249	.4191		.5409	
% OF EMP. COMMUTE BY PRIVATE AUTO	.9384				
% OF EMP. COMMUTE BY PUBLIC TRANSIT		.6558			
% OF EMP. COMMUTE BY WALKING			.8976		
% OF EMP. COMMUTE BY OTHER MEANS					
% OF HH. WITH ACCESS TO 1 AUTO	.5001	.8188		.6122	
% OF OCC. H WITH 1 UNIT	.8821			.4202	
% OF OCC. H WITH 2 UNITS				.5044	.4934
% OF OCC. H WITH 3 OR MORE UNITS		.9472	.4141		
% OF OCC. H UNITS BUILT BEFORE 1939					.8212
% OF OCC. H UNITS BUILT 1940-1949				.8196	
% OF OCC. H UNITS BUILT 1950-1959				.6975	
% OF OCC. H UNITS BUILT 1960-1969	.7065	.4631			
% OF OCC. H UNITS BUILT 1970-1979	.8008				
% OF OWNER OCCUPIED HOUSING UNITS		.9423			
% OF OCC. H UNITS DENSITY < THAN 1.01	.8127	.6953		.4741	
% OF OCC. H UNITS DENSITY > THAN 1.01		.4394	.4734		.5469
AGG. VALUE OF OWNER OCCP. NCO HOUSES	.7371			.4726	

1. The factor loadings area created by using a Principal Components Analysis promax rotation for the 23 variables chosen from previous factorial ecology and quality-of-life studies. These variables are tested separately to determine the amount of variance that is contributed to each factor and each variable's factor loading is listed. These 23 variables will be merged with the 57 variables used in the United Way data set to create the extended data set.

On the other hand, the five factors of the extended data set explained almost 80% of the variation. This result provides some justification for the extra effort involved in creating the extended data set.

For additional justification, I conducted a separate factor analysis to determine the amount of variance that is contributed by factor analyzing the 23 variables used in previous factorial ecology and quality-of-life studies. These are the variables that were added to the United Way's data base to create the extended data base (see Table 5.13). These 23 variables accounted for 21% of explained variance. If this amount is added to the 55% variance that the United Way's data set contributed, the total amount of explained variance is 76%. By examining the amount of variance that each data set contributes separately, I feel more confident that the increase of explanatory power which the extended data set provides in comparison to the United Way's data set is meaningful and not simply a artifact of the mathematical procedure itself.

Discussion of Table 5.15 Factor Loadings Within the Knoxville MSA

Factor one indicated a "white, blue collar, middle class family" dimension of the data set, as was the case previously (see page 70). The percentage of white population had a extremely high factor loading of .78. Five of the six variables measuring age had high factor loadings. The percentage of population less than five, between 10 and 19, and between 35 and 54 had factor loadings $\geq .67$ offering evidence of this factor's delimitation of families.

TABLE 5.15
STANDARDIZED FACTOR LOADINGS FROM THE EXTENDED DATA SET
FOR KNOXVILLE MSA CENSUS TRACTS USING
PRINCIPAL COMPONENTS ANALYSIS PROMAX ROTATION
BASED ON STANDARDIZED REGRESSION COEFFICIENTS

VARIABLES	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
% OF OCCUPIED HOUSING UNITS	.5540				
% OF URBAN POPULATION					.6604
% OF BLACK POPULATION					.8175
% OF WHITE POPULATION	.7821				
% OF INDIAN POPULATION					
% OF ASIAN POPULATION		.4537	.4109		
RATIO OF NWKRS TO WORKERS		-.4688	-.5845		.4485
% EMP. OUTSIDE COUNTY OR ST.	.6581	.5041			
% FEMALE HEADED HH WITH KIDS					.9304
% OF NONFAMILY HOUSEHOLDS		.4385		.5091	
% OF POP. GREATER THAN 65				.8584	
% OF POP. LESS THAN 5	.8691				
% OF POP. BETWEEN 10 AND 19	.6708		.7451		
% OF POP. BETWEEN 20 AND 34	.4252		.7882		
% OF POP. BETWEEN 35 AND 54	.8504				
% OF POP. BETWEEN 55 AND 64	.5266			.7171	
% OF POP. 25+ COMPLET. < 4YRS HS	.7417	-.5980			
% OF POP. 25+ COMPLET. 1 & 3 YEARS OF COLLEGE		.8962			
% OF POP. 25+ COMPLET. 4+ YRS OF COLLEGE		.9511			
% OF POP. WITH INC. < THAN \$5000					.4410
% OF POP. WITH INC. \$7,500 & \$9,999				.5486	
% OF POP. WITH INC. \$10,000 & \$14,999	.4298			.5717	
% OF POP. WITH INC. \$15,000 & \$24,999	.7271				
% OF POP. WITH INC. \$25,000 & \$49,999	.4942	.7361			
% OF POP. WITH INC. > THAN \$50,000		.8245			
% OF IND. WITH INC. < THAN \$5000			.8110		
% OF IND. WITH INC. \$7,500 & \$9,999			.7737		
% OF IND. WITH INC. \$10,000 & \$14,999		.6963		.4706	
% OF IND. WITH INC. \$15,000 & \$24,999		.8195			
% OF IND. WITH INC. \$25,000 & \$49,999		.8360			
% OF IND. WITH INC. > THAN \$50,000		.5333			
% EMPLOYED AS PRIVATE WAGE EARNERS	.4781	.4528			
% EMPLOYED IN STATE GOVERNMENT			.7581		
% EMPLOYED IN LOCAL GOVERNMENT	.6477				
% SELF EMPLOYED	.6006	.4485			

TABLE 5.15 (Continued)

VARIABLES	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
% EMP. IN AG., FORESTRY AND MINING	.5631				
% EMP. IN CONSTRUCTION	.8534				
% EMP. IN MANUFACTURING	.9082				
% EMP. IN TRANS-PUBLIC UTILITIES	.6863				
% EMP. IN RETAIL TRADE	.4960	.4895			
% EMP. IN F.I.R.E.		.8133			
% EMP. IN SERVICES		.7134			
% EMP. IN PUBLIC ADMINISTRATION	.4353	.5496			
% EMP. IN MANAGEMENT		.9236			
% EMP. IN TECH/SALES	.4812	.6932			
% EMP. IN SERVICE OCC.			.6347		
% EMP. IN PRECISION PROD.	.9191				
% EMP. IN OPERATOR-LABORER POSITIONS	.8758	.4988			
% OF OCCP. RENTAL HOUSING UNITS					.4854
% OF POP. 5+ IN SAME HOME AS IN 1975	.8437				
% OF POP. 5+ MOVE WITHIN STATE	.8647				.4116
% OF POP. 5+ MOVED FROM OTHER STATE		.5472	.4115		
% OF POP. 5+ ENTERED COUNTRY			.5810		
% OF POP. BELOW 75% OF POVERTY LEVEL		.5716			.4918
% OF POP. BETWEEN 75% & 125% OF POV.	.4833	.5033			
% OF POP. BETWEEN 125% & 150% OF POV.	.5010	.4852			
% OF POP. AT OR ABOVE 200% OF POV.	.6399	.6204			
% OF POP. OF SPANISH ORIGIN			.4034		
-----Extended Data Set-----					
% OF MARRIED COUPLE FAMILIES	.9182				
% OF POPULATION < 15 OR > THAN 65	.8493				
% OF UNEMPLOYED INCLUDED IN WORKFORCE			.9529		
% OF EMPLOYED POP. WORKING PART TIME			.9643		
% OF HH. ON PUBLIC ASSISTANCE				.7861	
% OF HH. ON SOCIAL SECURITY		.8463			
% OF EMP. COMMUTE BY PRIVATE AUTO	.8434	.5339			
% OF EMP. COMMUTE BY PUBLIC TRANSIT					.8889
% OF EMP. COMMUTE BY WALKING			1.0961 ¹		

1. It is possible to have factor loadings greater than 1, when promax solutions are utilized (Rummel, 1970)

TABLE 5.15(Continued) ¹.

VARIABLES	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
% OF EMP. COMMUTE BY OTHER MEANS					
% OF HH. WITH ACCESS TO 1 AUTO		.4242		.6271	
% OF OCC. H WITH 1 UNIT	.8539				
% OF OCC. H WITH 2 UNITS				.6553	
% OF OCC. H WITH 3 OR MORE UNITS	-.4765	.4654			.4252
% OF OCC. H UNITS BUILT BEFORE 1939		.5212		.5290	
% OF OCC. H UNITS BUILT 1940-1949				.8311	
% OF OCC. H UNITS BUILT 1950-1959				.6202	
% OF OCC. H UNITS BUILT 1960-1969	.4070	.4077			
% OF OCC. H UNITS BUILT 1970-1979	.5793	.4761		-.4365	
% OF OWNER OCCUPIED HOUSING UNITS					.4859
% OF OCC. H UNITS DENSITY < THAN 1.01		.7802			
% OF OCC. H UNITS DENSITY > THAN 1.01					.4469
AGG. VALUE OF OWNER OCCP. NCO HOUSES		.6162			

¹ The cluster results are based on factor loadings created by extracting 57 variables used in the United Way's inter-urban geo-demographic database system and 23 from previous factorial ecology and quality-of-life studies.

Three of the five income variables had high factor loadings. The loadings corresponded with income levels that would be expected of middle class households. Blue collar occupational variables had high factor loadings. The high loading on percentage of population greater than 25 completing less than four years of high school provides additional support for the blue collar interpretation of this group.

The factor loadings from the extended data set offer additional support of this interpretation. The factor loading on the percentage of married couple families is .92, which further enhances the familial interpretation of the dimension. The variable measuring single unit occupied housing had a significant loading providing further evidence of the familial characteristics. Factor loadings on age of occupied housing had significant loadings for the time periods between 1960-1969 and 1970-1979 indicating that these people were living in some of the most recently constructed housing in the area.

Factor two was further refined using the extended variable set compared to the same factor produced with the United Way variables (page 74). Using the United Way's data set, factor two indicated a "white collar singles and small families" dimension of the data set. An examination of the factor loadings from the variables added to create the extended data set indicates that this factor also seems to indicate a "white collar gentrification and nouveau riche" dimension of the data set. All three of the educational attainment variables have significant factor loadings on factor two.

The loadings measuring the percentage of population over 25 completing one to three years of college education and the percentage of population over 25 completing four or more years of college demonstrate the importance of higher education. Four of the five variables measuring individual income have significant factor loadings indicating either non-family households or high income dual income households. The factor loadings measuring white collar employment are significant, supporting the loadings of the education and income variables. The variable measuring the percentage of non-family households had a significant factor loading. However, importance of this variable may be overstated, because the two variables measuring family income above \$25,000 also have moderately high factor loadings on factor two.

The factor loadings from the extended data set provide additional information not previously uncovered about factor two. There is a significantly high loading on the percentage of households receiving social security income. People on social security, having few or no children at home, would have some characteristics in common with "empty nester" and non-family households. The variables measuring housing characteristics also embrace the expanded interpretation of the second factor. The housing of choice appears to be either older housing with distinctive architectural character or low maintenance newer housing in fashionable nouveau riche suburban areas. There are significant factor loadings for the variable measuring percentage of occupied housing built before 1939 and two variables measuring the percentage of occupied housing built after 1960.

The factor loading for occupied housing built before 1939 corresponds to the gentrification that has occurred in Knoxville, by persons who would load high on the second dimension. Gentrified neighborhoods normally consist of middle class singles, childless couples, and retirees (Hartshorn 1980). The significant factor loadings for the other housing variables show that this group also tends to live in luxurious modern housing, but tends to shy away from post WWII tract housing. Their housing preferences are reflected by their income. The significant loading on the percentage of occupied housing with three or more units identifies elderly and wealthy people living in apartments and condominiums, as well as white collar middle class families with a few or no children.

Factor three indicated an "university" dimension on the extended data set as it did with the United Way subset of variables (see page 74). There were high loadings on the variables that measured the percentage of Asian population, the percentage of Spanish population, and the percentage of the population entering the United States from other nations. These ethnicity variables may reflect people entering the area to study at the University. Variables measuring the percentage of population between 10 and 19 and the percentage of population between 20 and 34 had high factor loadings indicating the existence of teenage and young adult population, which would be expected in an university environment. The university dimension is further confirmed by examining employment variables. Employment variables with high factor loadings were the percentage of population employed in state government and percentage employed in service occupations.

Many students supplement their income with part time service positions, which was reflected in the high percentage of employed population working part time.

The factor loadings contributed by the extended data set provide additional evidence supporting the "university" interpretation of factor three. The percentage of unemployed included in total workforce may be students, or have similar education and employment characteristics that teens and young adults would have. The extremely high factor loading on the percentage of population working part time provides further support to the university dimension of the data set. The factor loading of 1.09 on the variable measuring employed commuting by walking would again reflect the university population.

There was a change, however, in the interpretation dimension of the data set indicated by factor four. With the United Way's data set, factor four represented an "elderly" dimension (page 75). I feel that the fourth factor using the extended data set indicates an "elderly, public assistance" dimension of the data set. There is a high factor loading for the percentage of population living in non-family households. The factor loadings for the variables measuring both family and individual income indicates a poor to lower middle class income level. Because of the significant factor loading for the two variables measuring family income, I conclude that poor, older families are captured by this dimension.

Factor loadings from the extended data set provide additional help in defining this factor.

The extremely high factor loading for the variable measuring the percentage of households receiving public assistance income defines the dependency characteristics of this group. The loadings of the variables measuring age of occupied housing units are high for the older units, which would indicate these people are "trapped" in units purchased long ago and from which they cannot afford to move.

Factor five indicated a "black, single parent" urban dimension of the data set for both the extended data set, as well as the United Way data set (page 75). I feel that there is a clearer poverty dimension using the extended data set than was elicited in the earlier analysis. There is a high factor loading for the percentage of black population. The high factor loading on urban population indicates that the distribution of black population is related to levels of urbanity. The high factor loading on female headed households with children is a characteristic of many black urban families.

With the extended data set, there was a significant factor loading for the percentage of population living below 75% of poverty level and for family income below \$5,000 which confirmed the poverty aspect of the areas scoring highly on this dimension. Neither of these variables had a significant factor loading when the United Way's data set was factor analyzed. With the extended data set, the factor loadings became significant, which represents an improvement in interpretability offered by the extended data set.

Further, the extended data set contributed significant factor loadings for the percentage of the employed commuting by public transit, the percentage of occupied housing units with three or more units, and the percentage of occupied housing units with density greater than 1.01 persons per room. All three of these variables would provide further support to the poverty aspect of the data set.

Step 3.3 Subjecting the Extended Data Set to Cluster Analysis

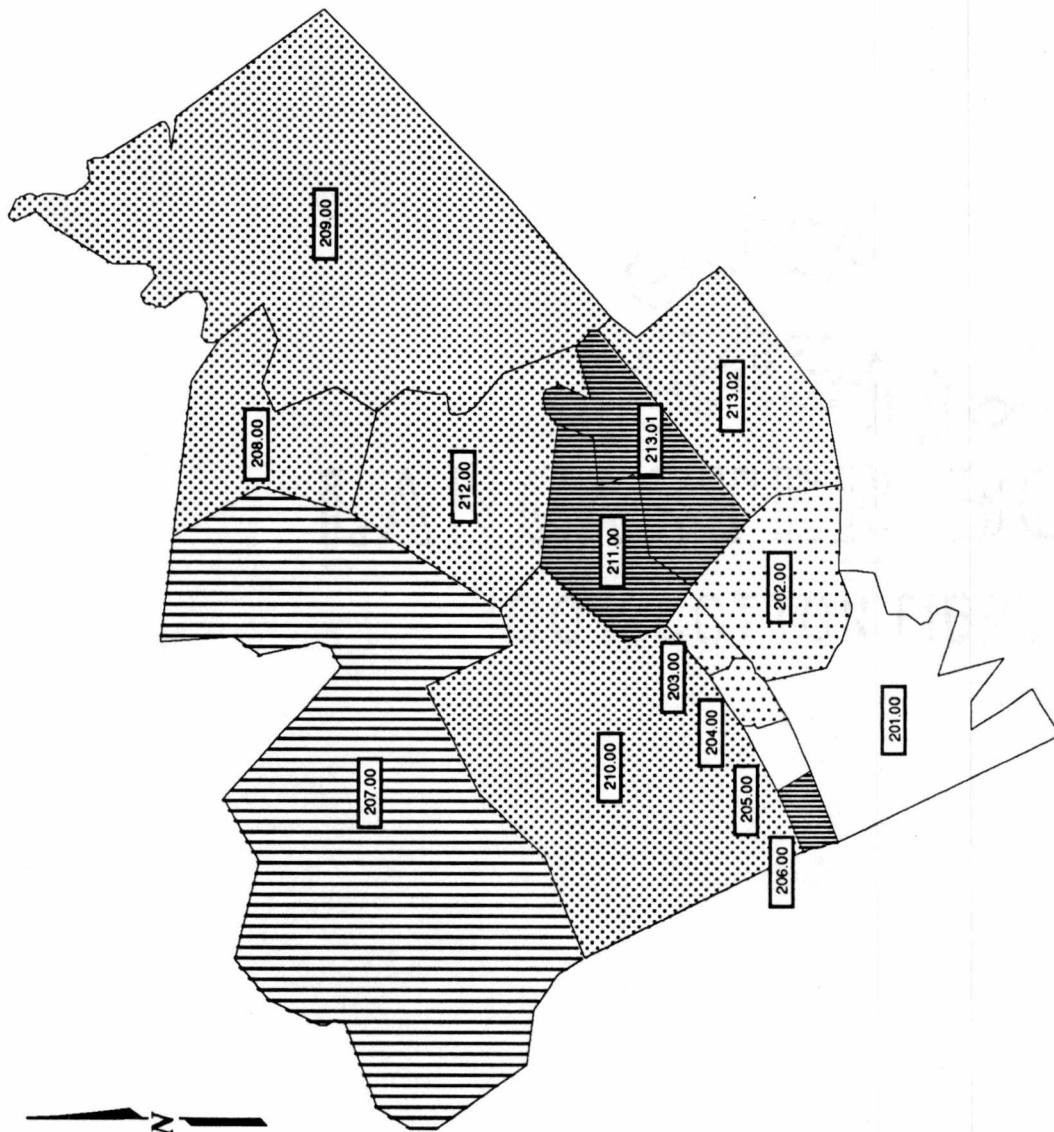
Using standardized factor scores from the extended data set, I used Ward's clustering method to group the 111 census tracts. I had previously determined Ward's method to be the best cluster analytic method from my research using the United Way's data set. Using the scree plot as a guide, I created an eleven cluster solution (see Figures 5.5 - 5.8, pages 108-118). To compare cluster assignment based on the United Way data set and the extended data set, see Table 5.16 (page 117). For a comparison of changes in census tract assignment, see Table 5.17 (page 118). Step 3.4 will compare cluster results between the United Way data set and the extended data set.

Step 3.4 Comparing Cluster Analysis Results

In order to compare cluster analysis results of the United Way's data set against those of the extended data set, I examined the number of clusters that each solution produced, the number of census tracts assigned to each cluster, and distribution of census tract assignments.

FIGURE 5.5
CLUSTER ASSIGNMENTS FOR ANDERSON COUNTY CENSUS TRACTS
USING EXTENDED DATA SET

CLUSTER ASSIGNMENTS FOR ANDERSON COUNTY CENSUS TRACTS USING EXTENDED DATA SET



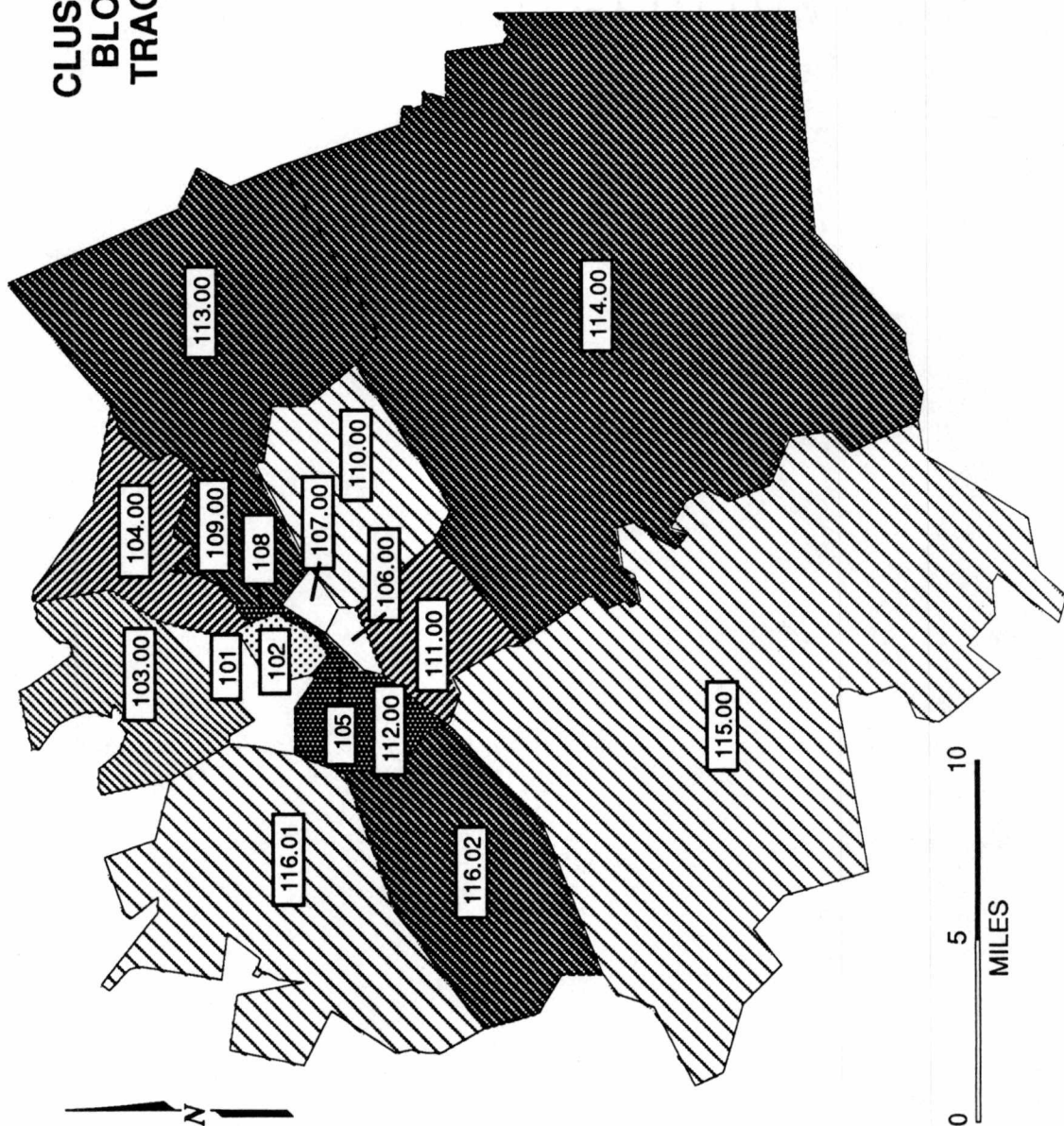
ASSIGNMENT

- | | |
|--|------------|
| | CLUSTER 1 |
| | CLUSTER 2 |
| | CLUSTER 3 |
| | CLUSTER 4 |
| | CLUSTER 5 |
| | CLUSTER 6 |
| | CLUSTER 7 |
| | CLUSTER 8 |
| | CLUSTER 9 |
| | CLUSTER 10 |
| | CLUSTER 11 |



FIGURE 5.6
CLUSTER ASSIGNMENTS FOR BLOUNT COUNTY CENSUS TRACTS
USING EXTENDED DATA SET

CLUSTER ASSIGNMENTS FOR BLOUNT COUNTY CENSUS TRACTS USING UNITED WAY DATA SET



ASSIGNMENTS

- CLUSTER 1
- CLUSTER 2
- CLUSTER 3
- CLUSTER 4
- CLUSTER 5
- CLUSTER 6
- CLUSTER 7
- CLUSTER 8
- CLUSTER 9

FIGURE 5.7
CLUSTER ASSIGNMENTS FOR KNOX COUNTY CENSUS TRACTS
USING EXTENDED DATA SET

CLUSTER ASSIGNMENTS FOR KNOX COUNTY CENSUS TRACTS USING EXTENDED DATA SET

ASSIGNMENTS

- CLUSTER 1
- CLUSTER 2
- CLUSTER 3
- CLUSTER 4
- CLUSTER 5
- CLUSTER 6
- CLUSTER 7
- CLUSTER 8
- CLUSTER 9
- CLUSTER 10
- CLUSTER 11
- NO DATA

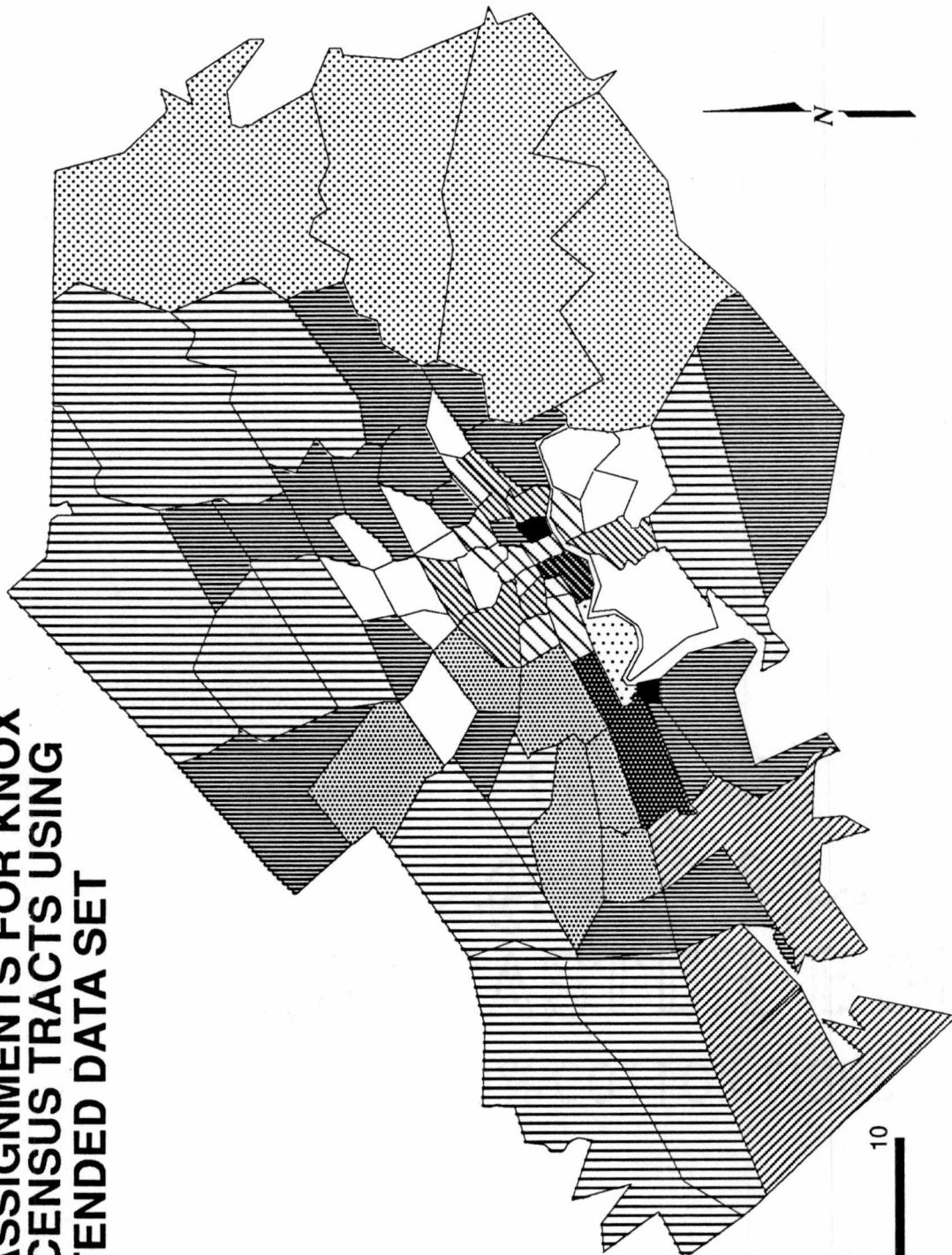


FIGURE 5.8
CLUSTER ASSIGNMENTS FOR UNION COUNTY CENSUS TRACTS
USING EXTENDED DATA SET

CLUSTER ASSIGNMENTS FOR UNION COUNTY CENSUS TRACTS USING EXTENDED DATA SET

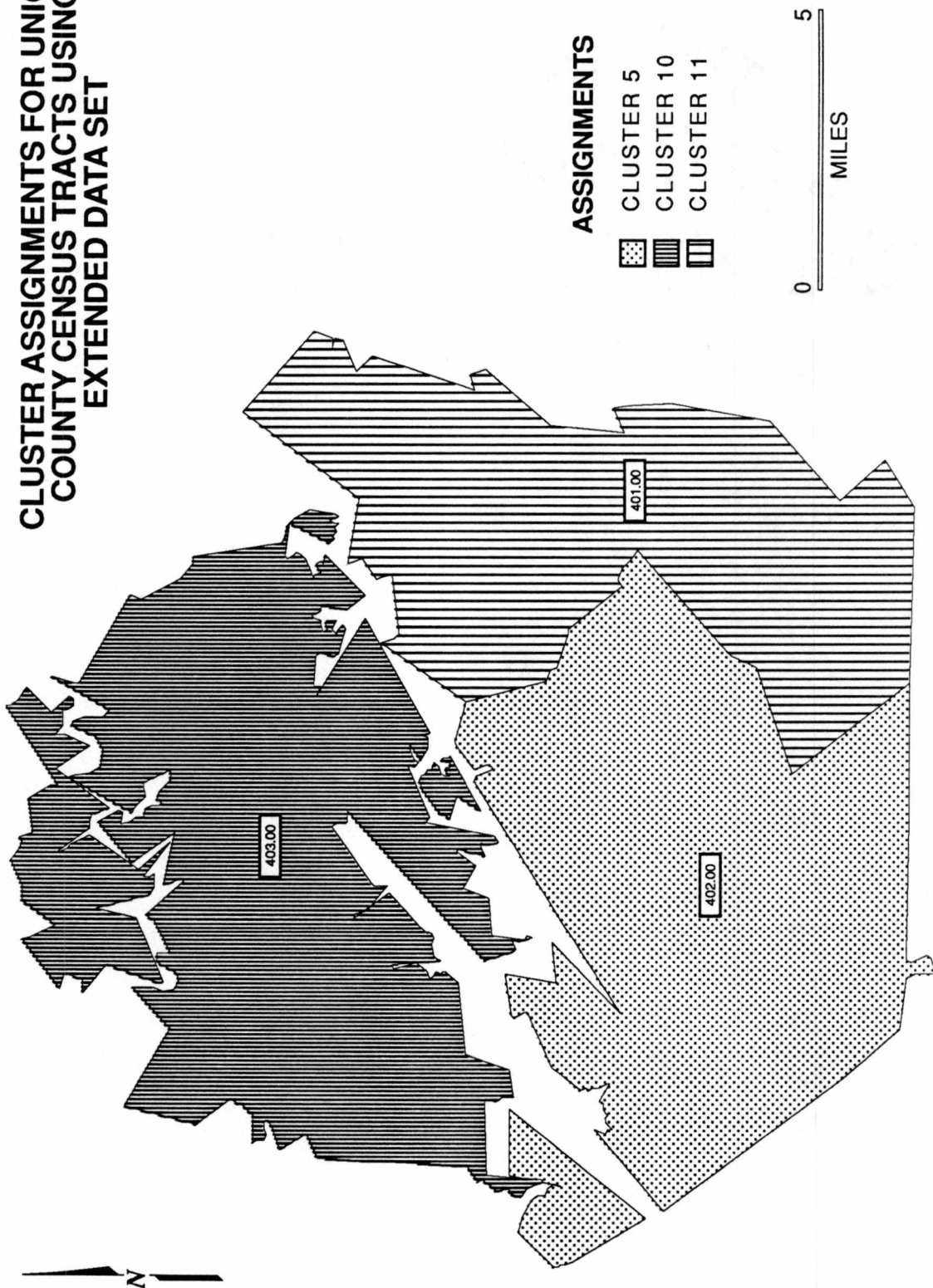


TABLE 5.16
CLUSTER RESULTS FOR KNOXVILLE MSA CENSUS TRACTS
COMPARING THE UNITED WAY'S DATA SET AND THE EXTENDED DATA SET

CLUSTER	NUMBER OF MEMBERS	CLUSTER	NUMBER OF MEMBERS
WARD'S CLUSTERING METHOD United Way Data Set ¹ .		WARD'S CLUSTERING METHOD Extended Data Set ² .	
1	18	1	18
2	5	2	4
3	14	3	12
4	10	4	8
5	13	5	12
6	13	6	7
7	13	7	3
8	11	8	2
9	14	9	2
		10	24
		11	19

1. The cluster results are based on the standardized factor scores created by extracting 57 variables used in the United Way's inter-urban geo-demographic database system for 111 of 113 census tracts.

2. The cluster results are based on standardized factor scores created by extracting 57 variables used in the United Way's inter-urban geo-demographic database system and 23 from previous factorial ecology and quality-of-life studies for 111 of 113 census tracts.

TABLE 5.17
COMPARISON OF NEW CLUSTER ASSIGNMENTS
BETWEEN THE UNITED WAY'S DATA SET ¹ AND THE EXTENDED DATA SET ².

OLD ASSIGNMENT		CENSUS TRACTS	NEW ASSIGNMENT
UNITED WAY DATA SET			EXTENDED DATA SET
CLUSTER 1	56.01		CLUSTER 11
CLUSTER 1	62.04		CLUSTER 11
CLUSTER 2	45		CLUSTER 6
CLUSTER 3	64		CLUSTER 11
CLUSTER 3	110		CLUSTER 11
CLUSTER 4	60		CLUSTER 11
CLUSTER 5	9		CLUSTER 8
CLUSTER 5	11		CLUSTER 8
CLUSTER 6 & 7	46.01		CLUSTER 11
CLUSTER 6 & 7	59.01		CLUSTER 11
CLUSTER 6 & 7	62.01		CLUSTER 11
CLUSTER 8	15		CLUSTER 1
CLUSTER 8	16		CLUSTER 1
CLUSTER 9	21		CLUSTER 10
CLUSTER 9	108		CLUSTER 11
CLUSTER 9	112		CLUSTER 11
NUMBER OF REASSIGNED AREAL UNITS			16
PERCENTAGE OF REASSIGNED AREAL UNITS			14.4%

1. The cluster results are based on standardized factor scores created by extracting 57 variables used in the United Way's inter-urban geo-demographic database system for 111 of 113 census tracts.

2. The cluster results are based on standardized factor scores created by extracting 57 variables used in the United Way's inter-urban geo-demographic database system and 23 from previous factorial ecology and quality-of-life studies for 111 of 113 census tracts.

Based on my examination of the results, I concluded that the extended data set provided a better solution. The five factor solution generated from the extended data set provided 79.5% explained variance. This result increases my confidence in the data set's ability to accurately depict the urban morphology of Knoxville.

Second, the four small clusters (i.e. those with fewer than four members) represent unique population groups within the overall data set. These lifestyle groupings would have been amalgamated into more heterogeneous clusters if only the subset of variables employed by the United Way had been used.

Step 3.5 Developing a geo-demographic database system

Based on the cluster results generated from the extended data set, the final step necessary to develop a geo-demographic data base system is to identify the clusters. I examined the standardized factor score means, standard deviations, and their distribution for each cluster see Table 5.18 (page 120). When examining Table 5.18 a negative mean score reflects that the population group represented by a particular factor has a lower than average representation within a particular cluster. Likewise, a positive mean score reflects that the population group represented by a particular factor has a higher than average representation within a particular cluster. A score near zero reflects that the population group represented by a particular factor is near the average for the study area.

TABLE 5.18
EXTENDED DATA SET'S CLUSTER MEAN'S AND STANDARD DEVIATIONS
FOR KNOXVILLE MSA CENSUS TRACTS ¹.

Cluster	Factor1	Factor2	Factor3	Factor4	Factor5
Cluster Mean's					
1	-0.1753	0.1988	-0.4644	0.6705	0.1773
2	-0.2819	1.6542	-0.1463	1.0863	-0.8425
3	-1.1326	-0.7873	-0.0693	0.0727	0.1466
4	-0.2396	-1.0130	0.2000	0.7274	2.0852
5	1.6240	-0.4279	0.2808	0.8985	0.0519
6	1.1437	1.8352	0.8190	0.0337	1.0080
7	0.9897	1.9655	-0.0293	-1.5204	-0.2935
8	-1.5423	0.1280	5.7744	-0.1712	0.8329
9	0.0139	2.5961	2.2407	2.2353	3.0368
10	-0.7238	-0.2074	-0.7783	-1.0826	-0.9409
11	0.5027	-0.3434	-0.3032	-0.3704	-0.5285
MAXIMUM	1.6240	2.5961	5.7744	2.2353	3.0368
MINIMUM	-1.5423	-1.0130	-0.7783	-1.5204	-0.9409
Cluster Standard Deviations					
1	0.1652	0.0863	0.0843	0.2129	0.0993
2	0.2225	0.1862	0.1122	0.8360	0.0623
3	0.5285	0.2556	0.4560	0.3606	0.3033
4	0.4691	0.3045	0.1585	0.4213	0.7583
5	0.4372	0.3688	0.3122	0.9252	0.3517
6	0.8054	0.8037	0.2936	0.5469	0.5961
7	0.3414	0.7305	0.3524	0.1430	0.1824
8	0.6028	0.5386	1.1394	1.7354	0.9695
9	0.8013	1.1294	0.0292	0.1710	0.0814
10	0.3746	0.5498	0.1833	0.3814	0.3223
11	0.5468	0.5889	0.1664	0.4580	0.2835

1. The source for this table was the standardized factor scores produced using the extended data set created using 57 variables from the United Way's inter-urban geo-demographic database system and 23 variables extracted from previous factorial ecology and quality of life studies.

For each cluster, I developed a label that would most appropriately describe the demographic and socio-economic characteristics of the population. For a listing of the extended data set's factor loadings, see Table 5.15 (pages 100-102).

Cluster 1 - Empty Nesters				
Factor1	Factor2	Factor3	Factor4	Factor5
-0.1753	0.1988	-0.4644	0.6705	0.1773

After examining the means of the standardized factor scores for cluster one, I felt that this cluster best represents the "empty nesters" population group. To make this determination, I examined the high mean score for factor four and the positive scores for factor two and factor five. As discussed previously, factor four represented the "elderly" dimension of the data set, factor two represented the white collar "gentrification" and "nouveau riche" dimension of the data set and factor five represented the poverty dimension of the data set. The mean factor scores provided by factor two and factor five that there is a higher than average concentration of population within this cluster that have characteristics reflected by factor two, factor four, and factor five. I felt that this group probably consists of an older, predominantly white collar, empty nester and single population nearing retirement or already retired. Considering the mean on factor five, there may be the existence of poverty pockets within an otherwise affluent population. My classification is supported by the negative mean scores for factor one, the blue collar family dimension and for factor three, the university population dimension.

A lower than average percentage of population within tracts assigned to this cluster have characteristics reflected by these latter two factors. Income for this population group would be moderate to low in some cases. The distribution of tracts assigned to cluster one corresponded to post-WWII suburban developments. Some examples of tracts exemplifying this group were tracts 22, 23 and 34 in south Knoxville.

Cluster 2 - Rich and Famous				
Factor1	Factor2	Factor3	Factor4	Factor5
-0.2819	1.6542	-0.1463	1.0863	-0.8425

With an high mean score for both factor two, the white collar "gentrification and nouveau riche" dimension and for factor four, the "elderly" dimension, I felt that this cluster defined the "rich and famous" population within Knoxville. The mean factor scores for factor two and factor four indicate that there is a significantly higher than average concentration of population with these characteristics.

This group consisted of a older, white collar executive and professional population, with some small families and empty nesters. There were four census tracts for this group. The interpretation of the cluster is further supported by the negative mean scores for factor one, the "blue collar family" dimension, factor three, the "university" dimension, and a high negative score for factor five, the "poverty" dimension. This is expected because there would be a below average concentration of population possessing the characteristics indicated by factor one, three and five in tracts assigned to this cluster. Tract 25, Sequoyah Hills, is assigned to this cluster.

This area is the oldest exclusive neighborhood in Knoxville and it contains many large and beautiful homes. The other three tracts; 201, 203, and 204, are in Anderson County. These are also wealthy areas where professionals employed at the facilities of the Oak Ridge National Laboratory as well as other high technology employers located in Oak Ridge and along the Pellissippi Parkway in west Knox County.

Cluster 3 - City Folks				
Factor1	Factor2	Factor3	Factor4	Factor5
-1.1326	-0.7873	-0.0693	0.0727	0.1466

For cluster three, the negative mean scores for factors one and two were more significant than the other scores. The negative scores for factor one and factor two indicate that the percentage of both blue collar and white collar middle class population with families would be significantly below average for this cluster. The slightly above average mean factor scores for factor four and factor five and the near average score for factor three are indications that this group consisted of lower middle class singles, female headed households with children, and elderly living in older, formerly blue collar middle class neighborhoods. Hence the designation as a "city folks" population group. Housing values are low, because these homes are extremely old. Further, within these census tracts, there are commercial and industrial land uses present which lower residential land and rental values. Tract one, the central business district, is included in this cluster. Another example of a tract assigned to cluster three was tract 26; the older Marble City neighborhood.

Cluster 4 - Urban Slums				
Factor1	Factor2	Factor3	Factor4	Factor5
-0.2396	-1.0130	0.2000	0.7274	2.0852

With a high mean score for factor five, and positive scores for factor three and four, I felt that this cluster best represented the urban slum population within the study area. Cluster four differs from cluster three because of the dramatic increase in the mean score on factor 5 which reflects the increase in population with "poverty" characteristics. This population would consist of singles, female-headed households, university population, and the elderly, which was similiar to cluster three's population. Income for this population would be the lowest of all clusters, reflecting the existence of subsidized housing projects and public assistance income. Housing is of extremely poor quality. Unique to this population group was the significant mean score on factor three, the university factor. This may be partially explained by the fact that female-headed households would have similiar age and income characteristics as the university population, as well as the existence of Knoxville College, a small, historically black college located near a public housing project in tract 14. That university's population would have similiar socio-economic characteristics as other impoverished people living in these tracts. Again, the mean scores for factor one and factor two are negative indicating that there is a lower than average concentration of population with characteristics indicated by these factors.

Cluster 5 - Miners and Farmers				
Factor1	Factor2	Factor3	Factor4	Factor5
1.6240	-0.4278	0.2808	0.8985	0.0519

With an above average mean factor score for factor four and a extremely high mean factor score for factor one, and positive scores for factor three and factor five, I felt that this cluster represented a primarily blue collar, rural middle class population with some elderly, university and poverty population. The negative mean score for factor two would be expected, because a white collar population seems out of place within this cluster. Tracts assigned to this cluster were found in east Knox County, southeast Union County, and eastern Anderson County. All three of these areas contain mining or quarrying activities, and farming. In eastern Knox County, limestone and zinc are mined. In eastern Anderson county, coal mining is prevalent. Farming employment is also important in these areas as well. There are many small and medium sized farms visible in the landscape. Tract 65, in east Knox County includes the zinc mining community of Mascot. It is an example of an area classified into cluster five.

Cluster 6 - New Frontier Suburbs				
Factor1	Factor2	Factor3	Factor4	Factor5
1.1437	1.8352	0.8190	0.0337	1.0080

Cluster six denoted "new frontier suburbs" These areas were developed beginning in the late 1950s, and continued into the 1970s. All five factors had mean scores greater than zero.

Factor two dominates with the highest mean factor score, showing the importance of white collar singles and small families. Factor one has the second highest mean factor score indicating a blue collar family population group. Surprisingly, factor five has a high mean score indicating that poor people also exist within this cluster. Factor three is also high indicating the existence of a university population group. With the exception of factor four, the "elderly" dimension, the other four factors are above average. This cluster would consist of both blue collar and white collar families with children, some singles, and female headed households with children. Again, as with most cluster's in the Knoxville MSA, the elderly exist but they are not as prevalent as in other groups. One example of a new frontier suburb was tract 45, the West Hills subdivision.

Cluster 7 - Exurban Families				
Factor1	Factor2	Factor3	Factor4	Factor5
0.9997	1.9655	-0.0293	-1.5204	-0.2935

Cluster seven delimited some of the newest areas of both white and blue collar populations. Irrespective of their occupational status, these populations were middle to upper middle class in economic standing and family orientation. The mean factor scores for factor one and factor two indicate there is an above average concentration of population with these characteristics. Factor two dominated this group. Factor one also had a significant mean factor score. The mean scores for factors three through five were negative indicating that these population groups are not prevalent.

From these mean factor scores, I concluded that this cluster reflected the characteristics of "exurban families". Cluster seven tracts are located in west Knox County, which is one area of recent residential growth. It consists of tracts 57.02, 58.01 and 58.02. Tracts 58.01 and 58.02 represent the Concord community. This area also has the beginnings of some retail and commercial development following the Kingston Pike arterial.

Cluster 8 - University				
Factor1	Factor2	Factor3	Factor4	Factor5
-1.5423	0.1280	5.7744	-0.1712	0.8329

It was evident from an examination of the mean factor scores that this cluster described the "university" population group. Factor three has an extremely high mean factor score. There is some significance to factor five, which would provide further evidence that these areas have low income, and the bulk of residents live in rental housing and dorms. The positive mean for factor two is perhaps indicative of full time university employees living near the campus as well as remnants of the white collar middle class population that once lived in these tracts. Negative mean scores for factor one and for factor four would be expected. In cluster eight, blue collar family and elderly populations would not be significant. It is not surprising that tract 9 and tract 11 were the only two tracts identified by this cluster. Tract 9 represents the site of the University of Tennessee, as well as a small, declining middle class neighborhood. Tract 11 represents the Fort Sanders area "student ghetto".

Again, along Kingston Pike, there was some strip commercial and retail activity primarily serving the needs of the university population.

Cluster 9 - Mixed Singles and Elderly
Factor1 Factor2 Factor3 Factor4 Factor5
0.0139 2.5961 2.2407 2.2353 3.0368

I had a difficult time describing cluster nine's characteristics. Tracts which formed the cluster had high mean scores for factors two through five, and a significant score for factor one. There are only two census tracts that are of this type, tract 37 and tract 44.01. Both of these tracts are located in Knoxville along Kingston Pike, west of Sequoyah Hills (tract 25). This is a zone in transition from residential and vacant into a commercial and retail development area. There are white collar middle class families, singles, university population, elderly and poor living in both single family homes and apartments. This was also an area that has experienced extreme growth in retail and commercial activity along Kingston Pike, which has changed its land use mix. With this in mind, I could only classify it as a mixed population.

Cluster 10 - Mature Neighborhoods
Factor1 Factor2 Factor3 Factor4 Factor5
-0.7238 -0.2074 -0.7783 -1.0826 -0.9408

The five mean factor scores gave little assistance in defining cluster ten's population significance.

All five of the mean factor scores were negative which reflects the fact that none of these individual factors pinpointed the positive characteristics of this population group. After studying the spatial distribution of the tracts assigned to this cluster, I decided that it reflects "mature neighborhoods". These neighborhoods are composed of an older, white collar middle class population. They were developed around the turn of the century and have retained their property value. Such areas are present within small cities in outlying counties. The distribution of this cluster follows Broadway north into Fountain City, and also into east Knoxville following Magnolia Avenue. One example of a tract assigned to the mature neighborhood group is Fountain City (tract 51). Other examples are tracts 211 and 213.01, the Lake City community in southeastern Anderson County.

Cluster 11 - Out to the Country				
Factor1	Factor2	Factor3	Factor4	Factor5
0.5027	-0.3434	-0.3032	-0.3704	-0.5285

Cluster eleven, "Out to the Country" had a slightly positive mean factor score for factor one, the blue collar family factor and negative mean factor scores for the other four factors. Considering the lack of development in tracts assigned to this cluster, I felt that this was best described as a rural, blue collar, lower middle class population. One example of cluster eleven was tract 59.01 Hardin Valley in west Knox County.

Discussion of Hypothesis II

I would have had a difficult problem in determining what population group characteristics typify each delimited cluster if the only information that I had about the study area were the cluster means and standard deviations. The standardized factor score means and standard deviations gave some indication of population characteristics, but without previous knowledge of the study area, my classification labels would have been more subjective. Even though a cluster has been assigned a particular designation, there are other population groups that co-exist or are unaccounted for. Cluster labeling is not very precise, because it is based more on my perception of the distribution of residential characteristics, than upon "hard" population and socio-economic data. After trying to label and explain cluster characteristics, I have increased my awareness of the difficulty researchers face in these daunting tasks.

On the basis of my research, I conclude that census tracts are too large to provide the desired homogeneity for a sound clustering system. One example is Tract 24, which consists of the Vestal community outside of the city of Knoxville, south of Fort Loudon Lake. This tract is assigned to cluster 4, "urban slums" which accurately reflects the characteristics of the majority of the tract population.

However, within this tract are two upper middle class housing developments; Cherokee Bluff Condominiums and Martha Washington Heights housing development. The population living in there latter housing developments were, in essence, improperly assigned to cluster 4, because of their location within Tract 24. From a target marketing perspective, the buying habits of people in these upper middle class housing developments would be significantly different than the average family living in Vestal. If a smaller areal unit was used, the population living within these upper middle housing developments might be separated from the lower income population living in Vestal.

Furthermore, the eleven cluster classification system that I developed using the extended data set provided some general information to differentiate population within the study area. Considering the uniqueness of each study area, one would have to develop his or her own system for each study ara. Most probably it would be necessary to conduct field research to verify the results.

The classification system developed for this thesis is based on now-dated 1980 Census information. Growth and development during the 1980s has changed the structure of the MSA. Unfortunately, 1990 data will not be readily available at the geographic scale necessary for geo-demographic research for another two years. Changes in morphological structure since the 1980 census are not reflected, and this shortcoming cannot be avoided.

The results indicate that I should accept my hypothesis that geo-demographic database systems would better classify population if the variables used to create "modern" geo-demographic database systems were

merged with variables used in classical urban ecological and quality-of-life studies. I created an extended data set to test my hypothesis using factor analysis, and then produced an eleven cluster classification system based on the standardized factor scores. With an almost 25% improvement in the amount of explained variance, the additional variables made it easier to identify data set dimensions, and to form more precise clusters. To verify the improvement, I conducted a separate factor analysis of the 23 variables that I added to the United Way system. This test provides additional support for the improvement engendered by using the extended system, in comparison to the United Way system.

The following chapter will summarize my research into geo-demographic database systems. I will review the major results of my research, discuss problems with current systems, and introduce areas of possible future research.

CHAPTER VI

CONCLUSION

The primary purpose of this thesis is to develop an intra-urban geo-demographic system. In this research the variables and classification methods used in the United Way of America's inter-urban geo-demographic system were applied at the intra-urban scale to the census tracts in the Knoxville, Tennessee MSA. By replicating the United Way system at a different geographic scale and testing my research hypotheses, I provided an introduction to geo-demographic database systems. To my knowledge, such systems have not been previously discussed in the geographic literature. Geo-demographic database systems segment the general population into relatively homogeneous categories by small areal units (e.g. census tracts, ZIP codes) or by community characteristics. For an intra-urban study, the spatial distribution of homogeneous areas within a city can be used to increase knowledge of urban morphology and neighborhood characteristics. For an inter-urban study, the community rating can give some indication of the socioeconomic and demographic composition of the overall population relative to other places. The four commercially produced geo-demographic database systems --PRIZM, ACORN, INFOMARK, and VISION--contain between 40 and 48 different categories to which to assign population. These categories are sufficiently detailed to identify most common population groups at both the inter-urban and intra-urban scale.

The primary use of geo-demographic database system data is by marketing researchers for trade area analysis and target marketing purposes. One example of an intra-urban application of geo-demographic database systems is by a seller who might desire to determine which population segment is most willing and best able to purchase his or her product. The census tracts which most closely match the demographic and socioeconomic background of the target consumer would be identified and sales information could then be mailed only to these persons.

Chapter I introduced two research hypotheses along with a stated agenda about how each hypothesis was to be tested. Hypothesis one examined the sensitivity of assignment of areal units to groups based on both the factor analytic and clustering techniques used. The results indicated that I should accept the hypothesis that group assignment was highly sensitive to both the type of factor analytic and clustering techniques utilized. Hypothesis two tested the sensitivity of cluster assignments to the input variables used. The results indicated that there would be a dramatic improvement in the amount of explained variance and the ease and precision of cluster definition if variables from previous factorial ecology/quality-of-life studies were merged with variables used to create an existing geo-demographic database system.

In Chapter II, I provided a brief introduction to commercial geo-demographic database systems. I reviewed the four systems currently available, and the United Way's inter-urban counterpart. I discussed some shortcomings which exist in geo-demographic database systems and explained the use of system output as one important type of input into even more elaborate desktop demographic systems.

Desktop demographic systems provide detailed information about consumers including their expenditure patterns and demographic characteristics. Considering the complexity of desktop demographic systems, I decided to focus solely on an examination of geo-demographic database systems, which are a major component in their development.

In Chapter III, I reviewed previous research in urban structural modeling and how previous research has provided the foundation for current geo-demographic database systems. I reviewed urban ecological models, social area analysis models, factorial ecology models, and quality-of-life studies. These models focus more on describing the process of morphological change than upon providing information that could be directly relevant to a characterization of the socioeconomic characteristics of the population at the census tract or ZIP code level. Many of the variables used in these earlier studies are, however, used to create geo-demographic database systems.

In Chapter IV, I reviewed factor analytic and clustering techniques commonly used to create geo-demographic database systems. First factor analysis and then cluster analysis are used to develop an areal unit classification system. Most commercial users of geo-demographic database systems are not familiar with these statistical techniques. I felt, therefore, that these procedures needed to be examined for their effect on the resultant classification system. I reviewed the most commonly used classification techniques, the interpretation of results, and problems associated with using these techniques.

Chapter V tests the two research hypotheses introduced in Chapter I.

Chapter V also contains information that can be used to develop a geo-demographic database system. I was fortunate in being able to obtain the list of variables used to create the United Way's system which greatly aided the first step -- database development of system design. In the second step of system design, I subjected the database to various factor analytic techniques and rotations. The technique which maximized explained variance without violating the factor analytic constraints was selected. Next, the significant factor loadings for each factor were examined to determine the appropriate dimensionality needed for areal discrimination. Standardized factor scores for census tracts were generated based on the factor loadings in order to serve as input into cluster analysis. Various clustering methods were employed and the best method selected. To test the second hypothesis, I created the extended data set by merging 23 additional variables from previous factorial ecology and quality-of-life studies with the existing United Way data set. This was done with an eye towards greater comprehensiveness and areal discrimination. Using Principal Components Analysis with Promax Rotation, I factor analyzed the extended data set. Using standardized factor scores, I clustered the extended data set using Ward's method, which had proven to be the best method for the United Way's data set considered separately. Then, the appropriate number of clusters were determined and the standardized factor scores were then used to calculate cluster means and standard deviations.

The last step in the process of system design is to label the clusters using information from the cluster means and standard deviations and the table of factor loadings as guides.

Problems With Using Geo-Demographic Database Systems

The results of my research indicate that there are some problems that geo-demographic database system users need to be concerned with when interpreting output. One problem is the appropriate unit of analysis. All of the research that I conducted employed data at the census tract level. My analytical results indicate that the census tract is too large to accurately segment the population into homogeneous groups. Geo-demographic database systems using census tracts as their basic areal unit provide much information about neighborhood composition and urban morphology. But, they also may miss significant population groups within a census tract which differ dramatically from the average tract values. There were, for example, many divergent population groups which existed within most of the Knoxville MSA census tracts that were essentially ignored by even the classification system based on the extended data set. I do not feel ZIP code data would have been more helpful, because these areas are normally larger than census tracts and, therefore, even less homogeneous. This data will be available in "Topologically Integrated Geographic Encoding and Referencing system format" (TIGER) format. Developed by the Census Bureau for the 1990 census, TIGER coverage has increased from only 16% of the United States in 1980 to almost 100% coverage in 1990. By digitizing the entire street grid of the United States, the information in the TIGER file represents an dramatic improvement for users of Census data.

A copy of **TIGER** for the entire United States is currently available on magnetic tape at the cost of \$87,000. By the end of 1990, **TIGER** should be available on CD-ROM. Each disk would cost \$250 and the set of 40 disks offering coverage of the entire United States would cost \$10,000, a modest investment for a firm adept at making good use of such detailed information (Schwartz 1990a; Larson 1990).

Digitization of public data has reduced costs associated with data access and has increased the number of potential sources and amount of data that are available to the researcher. Census 90 information should be released sometime in early 1991, after congressional redistricting occurs. Most census data will be released in a format that is compatible with current GIS software. It is expected that state and local agencies will follow suit, providing more public domain data for the researcher than ever before (Schwartz, 1990b)

Another way that this data heterogeneity problem could be ameliorated is if the ZIP+4 areal unit was introduced. This would reduce the number of people from 2,500 to 8,000 at the census tract level to somewhere between 4 and 80 households at the ZIP+4 level (U.S. Bureau of the Census 1983a; Riche 1990).

The ZIP+4 areal unit would represent a significant improvement in the usefulness of geo-demographic database system output, because these areas would be smaller and more homogeneous than even the block group (Riche 1990).

Another problem with geo-demographic database systems is data set development. Most systems are based on the input of a large number of census variables.

Considering the large number of variables, access to census tapes is essential. Yet this may prove to be a difficult and time consuming process for the user unfamiliar with the nuances of the job control language (JCL) needed to manipulate large data sets. Apparently, Census of Population and Housing data for 1990 will be released in TIGER format on CD-ROM. It is hoped that software designed to access 1990 census data will be more user-friendly than the software developed to access previous census data.

The inherent subjectivity of both factor analysis and cluster analysis is another problem of which the user must be aware. There are many factoring methods and rotations available. Each produces its own unique solution. There are some rules of thumb to help determine the proper number of factors and to identify the underlying dimensionality of the data set. But, in the end, the researcher has to decide which method best classifies the data set for his or her purposes. Likewise, cluster analysis is also highly subjective. There are some statistics that are useful in helping to determine the optimal number of clusters, but, in the end, it is up to the researcher to determine how many clusters are needed to properly classify the data set. The subjectivity of these methods cannot be avoided, but the user needs to be aware of the potential interpretation problems which might arise.

Implications For Users

There are three areas of concern that users need to be aware of when using geo-demographic database systems.

First, the purpose of these systems is to provide a method based on statistical techniques which can be used to segment the population into groups and assign them to unique clusters reflecting their socioeconomic and demographic characteristics. However, peak performance of geo-demographic demographic systems is prevented because of the problems associated with the census tract areal unit.

Second, for the user who is experienced using computers and has a basic understanding of statistics, the research presented in this thesis could be used as a guide to develop his or her own customized system. One could develop a personalized database using the extended data set as an example, and add additional variables if desired. For interpretive purposes, it is important for users to know which variables were included to create the lifestyle clusters. Commercial geo-demographic database systems vendors are unwilling to release information about the variables and methodology used to develop their systems due to the proprietary nature of their business. If a business person or academic researcher appreciates the advantages afforded by a geo-demographic database system output and cannot afford to obtain one of the commercial systems, it is feasible to develop one's own system.

Third, there is a tradeoff between the ability of the statistical techniques (factor analysis and cluster analysis) to classify the number of population groups and the marketer of commercial geo-demographic database system's need to subdivide the population. From the academic perspective, there will always be population groups with such diverse characteristics that they should not be assigned to an existing cluster, because they contribute a unique understanding to urban morphology.

However, commercial marketers of geo-demographic database systems claim that their systems can classify any population group at any level of geographic detail (e.g. census tract, ZIP code). This claim is incorrect because the nuances of factor analysis and cluster analysis will never permit every population subgroup to be properly classified. The user must understand that these systems provide general information and are one of many sources that should be reviewed when marketing decisions are being made.

Areas of Future Research

One area of future research would be to replicate the Knoxville intra-urban study in other cities. Knoxville's matrix of factor loadings could be compared with the matrices for the other cities using Target rotation. Target rotation would use the Knoxville matrix as the target to be compared with matrices derived from the other cities. In this way, the degree to which the results obtained from Knoxville are unique or symptomatic of broader urban morphological processes might be better assessed.

Target rotation is the only rotational scheme that uses unrotated factor loadings. An advantage to this method is that there is no induced manipulation of the factor loadings. The problem with this rotational method is the same as the problems associated with unrotated factor loadings in general: they are harder to interpret, do not achieve simple structure, and are harder to define (Rummel 1970; Weiss 1971).

Another area of future research would be to replicate this study using the ZIP+4 areal unit for the Knoxville MSA. I feel that the performance of the system would improve significantly using both 1990 data and a smaller, more homogeneous areal unit.

I would be interested in determining whether the same five factor Principal Components Analysis Promax Rotation solution would prove to be the best factor analytic method and rotational schema, as was the case using census tract level information.

Further, I would be interested to see if an eleven cluster Ward's method solution would again prove to be the best clustering solution. My hypothesis would be that this new system would produce more significant and differentiated clusters to identify the population groups than did the extended system based on census tracts.

A third possible area of future research would be to concentrate on adding variables that will detect industrial and commercial land uses. Current geo-demographic database systems are designed to classify population by the residents' demographic and socioeconomic characteristics, ignoring commercial and industrial land uses within the city. My experience in using the extended data set in the Knoxville MSA was that census tracts having mixed land uses were more difficult to classify properly. My recommendation would be to collect land use data for the appropriate scale of analysis. Land uses could be divided into several categories to encompass most land use within cities and these categories treated as variables in an extended geo-demographic database system.

REFERENCES

REFERENCES

- Aldenderfer, Mark S., and Roger K. Blashfield. Cluster Analysis. Beverly Hills, Sage, 1984.
- Anderberg, Michael R. Cluster Analysis for Applications. New York: Academic Press, 1973.
- Bederman, Sanford H. "The Stratification of 'Quality of Life' in the Black Community of Atlanta, Georgia." Southeastern Geographer 14 (1974) 1:26-37.
- Berry, Brian J. L. and Frank E. Horton. Geographic Perspectives on Urban Systems. Englewood Cliffs, N.J.,: Prentice Hall, 1970
- Berry, Brian. J. L. Editor. City Classification Handbook: Methods and Applications. New York:Wiley, 1972.
- Berry, Brian J. L., and Robert J. Tennant. Commercial Structure. Chicago: Northeast Illinois Planning Commission as cited by Berry, Brian J. and Frank E. Horton. 1970. Geographic Perspectives on Urban Systems. Englewood Cliffs, Prentice Hall, 1970.
- Blashfield, Roger K. and Mark S. Aldenderfer. "The Literature on Cluster Analysis." Multivariate Behavioral Research. 13(1978)271-295.
- Borgen, Fred H. and David M. Weiss. "Cluster Analysis and Counseling Research." Journal of Counseling Psychology. 18(1971) 6:583-591.
- Bourne, Larry S. "Physical Adjustment Processes and Land Use Succession: A Conceptual Review." Economic Geography. 47(1971) 1:1-15.
- Claritas. "PRIZM: The Standard in Geo-Demographic Targeting A Pocket Guide to the Five Micro-Geographic Systems." Claritas, 1985 (Advertisement)
- Crispell, Diane. "Locating Customers." American Demographics. 10(1988) 2:56.
- Crispell, Diane. "CACI Conquers Geography." American Demographics. 10(1988) 10:56.
- Crispell, Diane. "Finding Direction." American Demographics. 11(1989) 1:70.
- Curry, David J. "Single Source Systems: Retail Management Present and Future." Journal of Retailing. 65(1989) 1:1-21.

Cutter, Susan L. Rating Places: A Geographer's View on Quality of Life. Washington, D.C.: Association of American Geographers. Resource Paper Number 4. 1985

Dickinson, Joshua C. III, Robert J. Gray, and David M. Smith. "The 'Quality of Life' of Gainesville Florida: An Application of Territorial Social Indicators." Southeastern Geographer. 12(1972) 2:121-132.

Donnelly Marketing Information Services. CLUSTERPLUS Fact Book. Donnelly Marketing Information Services. April, 1988

Edmonson, Brad. "America By The Numbers: How Well Do ZIP Code Clusters Match Reality." American Demographics 11(1989) 8:45-46.

Elgie, Robert A. and Alex Rees Clark. "The Metropolitan Structure of Southern Metropolitan Areas: A Comparative Analysis." Urban Geography. 1(1980) 4:317-334.

Francese, Peter K. "Perspective." American Demographics. 10(1990) 1:17-18.

Harman, Harry H. Modern Factor Analysis. Second Edition. Chicago: The University of Chicago Press. 1967.

Hartshorn, Truman A. Interpreting the City: An Urban Geography. New York, John Wiley and Sons, 1980.

Hoyt, Homer. The Structure and Growth of American Cities. Washington, D.C.: The U.S. Government Printing Office, 1939.

Johnston, Ronald John. Multivariate Statistical Analysis in Geography New York:Longman, Inc, 1980.

Kalish, David. "A Shoppers Guide: Different Keystrokes for Different Media Folks are Offered by a Variety of Software Systems." Marketing and Media Decisions. 24(1989) 10:61-62, 64-65.

Kirchner, Russell J. and Richard K. Thomas. "Desktop Demographics." American Demographics. 11(1989) 8:4,35-37.

Larson, Jan. "Inside The Information Industry: TIGER Opens New Mapping Vistas for Businesses." American Demographics 12(1990) 6:16,18.

Liu, Ben-Chieh. Quality of Life Indicators in U.S. Metropolitan Areas. New York: Praeger, 1976.

Lorr, Maurice. Cluster Analysis for Social Scientists. San Francisco: Jossey-Bass Publishers, 1983.

Mason, Janet. "Dun & Bradstreet Creates Advantage from Data Diversity." ComputerWorld Second Edition. Volume 23 Sept. 11, 1989. Pages 78-79.

McHarg, Ian L. Design With Nature. New York: The Natural History Press, 1969.

Morsch, William and William F. Shaw. "Use of ZIP Codes in Location Analysis." The Appraisal Journal. 55(1987) 1:69-78.

Murdie, Robert A. Factorial Ecology of Metropolitan Toronto 1951-1961. Chicago: Department of Geography Research Paper #116, 1971.

National Decision Systems. "We Reached Beyond Our Vision To Bring You Single Source Marketing." American Demographics. 9(1987) 1:32-33. (Advertisement)

National Decision Systems. "Know Your Customer, Market, Sales Potential, Location," American Demographics. 12(1990) 1:32-34. (Advertisement)

National Decision Systems. "Vision Targeting Marketing System," National Decision Systems. December, 1989 (Advertisement)

O'Hare, Barbara Clark. "How to Profit from Clusters." American Demographics. 10(1988) 42-43.

O'Hare, Barbara. "1986 Comparative Cities Project Factor and Cluster Analyses Final Report." Unpublished United Way Report., 1986.

O'Hare, Barbara. Phone Interview 06-26-90

Park, Robert E., Burgess, Ernest W. and Roderick D. McKenzie. The City. Chicago: University of Chicago Press, 1925.

PC Week. "Extra-Crispy Demographics." March 20, 1987. 9(1987) 11:7.

Pols, Cynthia. "Bill Tailors ZIP Codes to City Needs." Nation's Cities Weekly. 12(1989) 21:8.

Punj, Girish and David W. Stewart. "Cluster Analysis in Marketing Research: Review and Suggestions for Applications." Journal of Marketing Research. 30(1983) 5:134-148.

Rees, Phillip H. The Factorial Ecology of Metropolitan Chicago. Masters Thesis, Department of Geography, University of Chicago, 1968.

Riche, Martha Farnsworth. "New Frontiers for Geodemographics." American Demographics. 12(1990) 6:20

Rosenblatt, S.M. "Empirical Study of the Sampling Error of Principal Components and Varimax Rotation Factor Loadings. Proceedings of the 77th Annual Convention of the American Psychological Association. 4(1969) 115-116.

Rummel, R. J. "Understanding Factor Analysis." Journal of Conflict Resolution. 11(1967) 4:24-59.

Rummel, R. J. Applied Factor Analysis. Evanston: Northwestern University Press, 1970.

SAS Institute Inc. SAS User's Guide: Statistics. Version 5 Edition. Cary N.C., 1985.

Schwartz, Joe. "Inside the Information Industry: Who Owns Who, and What It Means to You." American Demographics. 12(1990) 1:14,16-17.

Schwartz, Joe. "Perspective." American Demographics 12(1990a) 6:19.

Schwartz, Joe. "Perspective." American Demographics. 12(1990b) 8:18.

Shoultz, Donald. "PC Generated Data Pinpoints Markets." American Banker. 153(1988) 9:6-7.

Shevky, Eshref and Wendell Bell. Social Area Analysis. Stanford California: Stanford University Press, 1955.

Shevky, Eshref and Marilyn Williams. The Social Areas of Los Angeles: Analysis and Typology Los Angeles: The University of California Press, 1949.

Smith, David M. Geography Of Social Well Being In the United States. New York: McGraw Hill, 1973.

Sokal, Robert R. and Peter H. A. Sneath. Principles of Numerical Taxonomy. San Francisco: W. H. Freeman and Company, 1963.

Stebbins, Andy. "1990 Census: Boom or Bust for Marketers." Advertising Age. 58(1987) 9:18,20.

Stevenson, Dick. "Research and Reality: The Clustering of America." Marketing and Media Decisions. 24(1989) 4:124.

Taylor, Peter J. 1977. Quantative Methods in Geography: An Introduction to Spatial Analysis. Boston: Houghton Mifflin, 1977

Toomey, Terry J. Correspondence from National Planning Data Corporation., 1990.

U.S. Bureau of the Census. Geographic Information Code Scheme.
Volume PHC80-R5. Washington DC: U.S. Government Printing Office. 1983a.

U.S. Bureau of the Census. Knoxville, Tennessee: Census of
Population and Housing, 1980. Volume PHC80-2-0203.
Washington DC: U.S. Government Printing Office, 1983b.

Weiss, David J. "Factor Analysis and Counseling Research." Journal
of Counseling Psychology. 17(1970) 5:477-485.

Weiss, David J. "Further Considerations in Applications of Factor
Analysis." Journal of Counseling Psychology. 18(1971) 1:85-92

Weiss, Michael J. The Clustering of America. New York: Harper
and Row, 1988

Wood, Wally. "Conquest: A Hands-On Appraisal." Marketing and Media
Decisions. 23(1988) 6:148-149

APPENDIXES

APPENDIX 1 - UNITED WAY DATA SET

Variable Label	Variable Description
POCHU	% OF OCCUPIED HOUSING UNITS
PURBA	% OF URBAN POPULATION
BLAPO	% OF BLACK POPULATION
WHAPO	% OF WHITE POPULATION
INDPO	% OF AMERICAN INDIAN, ESKIMO AND ALEUTIAN POPULATION
ASIPO	% OF ASIAN, PACIFIC ISLANDER POPULATION
DEPRO	% NONWORKING PERSONS AS A RATIO TO EMPLOYED PERSONS OVER 16
WICOS	% OF EMPLOYED PERSONS 16+ WORKING OUTSIDE OF COUNTY OR STATE OF RESIDENCE
FEMHH	% OF FEMALE HEADED HOUSEHOLDS
NFAHH	% OF NONFAMILY HOUSEHOLDS
POP65	% OF POPULATION 65 AND OLDER
POPL5	% OF POPULATION LESS THAN 5
P1019	% OF POPULATION BETWEEN 10 AND 19
P2034	% OF POPULATION BETWEEN 20 AND 34
P3554	% OF POPULATION BETWEEN 35 AND 54
P5564	% OF POPULATION BETWEEN 55 AND 64
L4YHS	% OF POPULATION 25+ WITH LESS THAN 4 YEARS OF HIGH SCHOOL
Y13CO	% OF POPULATION 25+ WITH BETWEEN 1 AND 3 YEARS OF COLLEGE
Y4PCO	% OF POPULATION 25+ WITH 4 OR MORE YEARS OF COLLEGE
HINP1	% OF HOUSEHOLDS WITH YEARLY INCOME LESS THAN \$5000
HINP3	% OF HOUSEHOLDS WITH YEARLY INCOME BETWEEN \$7,500 AND \$9,999
HINP4	% OF HOUSEHOLDS WITH YEARLY INCOME BETWEEN \$10,000 AND \$14,999
HINP5	% OF HOUSEHOLDS WITH YEARLY INCOME BETWEEN \$15,000 AND \$24,999
HINP6	% OF HOUSEHOLDS WITH YEARLY INCOME BETWEEN \$25,000 AND \$49,999
HINP7	% OF HOUSEHOLDS WITH YEARLY INCOME GREATER THAN \$50,000

APPENDIX 1 (Continued)

Variable Label	Variable Description
PINC1	% OF UNRELATED PEOPLE 15 + WITH PERSONAL INCOME LESS THAN \$5,000 PER YEAR
PINC2	% OF UNRELATED PEOPLE 15 + WITH PERSONAL INCOME BETWEEN \$5,000 AND \$7,999 PER YEAR
PINC4	% OF UNRELATED PEOPLE 15 + WITH PERSONAL INCOME BETWEEN \$10,000 AND \$14,999 PER YEAR
PINC5	% OF UNRELATED PEOPLE 15 + WITH PERSONAL INCOME BETWEEN \$15,000 AND \$24,999 PER YEAR
PINC6	% OF UNRELATED PEOPLE 15 + WITH PERSONAL INCOME BETWEEN \$25,000 AND \$50,000 PER YEAR
PINC7	% OF UNRELATED PEOPLE 15 + WITH PERSONAL INCOME GREATER THAN \$50,000 PER YEAR
PEMPW	% OF WORKERS OVER 16 EMPLOYED AS PRIVATE WAGE EARNERS
PEMSG	% OF WORKERS OVER 16 EMPLOYED AS STATE GOVERNMENT WORKERS
PEMLG	% OF WORKERS OVER 16 EMPLOYED AS LOCAL GOVERNMENT WORKERS
PSEEM	% OF WORKERS OVER 16 SELF EMPLOYED
PEMS1	% OF WORKERS OVER 16 EMPLOYED IN AGRICULTURE, FORESTRY AND FISHING
PEMS2	% OF WORKERS OVER 16 EMPLOYED IN CONSTRUCTION
PEMS3	% OF WORKERS OVER 16 EMPLOYED IN MANUFACTURING
PEMS4	% OF WORKERS OVER 16 EMPLOYED IN TRANSPORTATION AND PUBLIC UTILITIES
PEMS6	% OF WORKERS OVER 16 EMPLOYED IN RETAIL TRADE
PEMS7	% OF WORKERS OVER 16 EMPLOYED IN FINANCE, INSURANCE AND REAL ESTATE
PEMS8	% OF WORKERS OVER 16 EMPLOYED IN SERVICES
PEMS9	% OF WORKERS OVER 16 EMPLOYED IN PUBLIC ADMINISTRATION

APPENDIX 1 (Continued)

Variable Label	Variable Description
PEM01	% OF WORKERS OVER 16 EMPLOYED IN MANAGEMENT POSITIONS
PEM02	% OF WORKERS OVER 16 EMPLOYED IN TECHNICAL, SALES, AND ADMINISTRATION POSITIONS
PEM03	% OF WORKERS OVER 16 EMPLOYED IN SERVICE POSITIONS
PEM05	% OF WORKERS OVER 16 EMPLOYED IN PRECISION PRODUCTION POSITIONS
PEM06	% OF WORKERS OVER 16 EMPLOYED IN OPERATORS, FABRICATORS, AND LABORER POSITIONS
PRHUN	% OF OCCUPIED RENTAL HOUSING UNITS, 1980
PNOTM	% OF POPULATION OVER 5 LIVING IN SAME HOME AS IN 1975
PINST	% OF POPULATION OVER 5 LIVING IN SAME STATE AS IN 1975
PWICO	% OF POPULATION OVER 5 THAT MOVED FROM ANOTHER STATE SINCE 1975
PFRAB	% OF POPULATION OVER 5 THAT MOVED FROM ANOTHER COUNTRY SINCE 1975
POVP1	% OF POPULATION WITH 1979 INCOME LESS THAN 75% OF POVERTY LEVEL
POVP2	% OF POPULATION WITH 1979 INCOME BETWEEN 75% AND 125% OF POVERTY LEVEL
POVP3	% OF POPULATION WITH 1979 INCOME BETWEEN 126% AND 149% OF POVERTY LEVEL
POVP4	% OF POPULATION WITH 1979 INCOME GREATER THAN 200% OF POVERTY LEVEL
POPSO	% OF POPULATION WITH SPANISH ORIGIN

APPENDIX 1 (Continued)

United Way Variables That Could Not Be Used in the Present Analysis ¹.

General Revenue of Local Government, 1977
Debt Ratio (per 1\$ of Revenue)
Debt Per Capita (Local Government)
Direct Expenditures Per Capita (Local Government)
Taxes Per Capita (Local Government)
Change in Population 1980-1985
Household E.B.I., 1985

Source: O'Hare, 1986.

¹. These variables were unused because they are unavailable at the census tract level. The United Way's geo-demographic data set was developed for a inter-urban study. I extracted the variables that were available at the census tract level for my intra-urban study of Knoxville.

APPENDIX 2 - VARIABLES USED TO EXTEND THE UNITED WAY SYSTEM

Variable Label	Variable Description
POMCF	% OF MARRIED COUPLE FAMILIES
P1565	% OF POPULATION UNDER 15 AND OVER 65
PUNEM	% OF UNEMPLOYED POPULATION INCLUDED IN TOTAL WORKFORCE
PWL40	% OF WORKERS EMPLOYED LESS THAN 40 HOURS PER WEEK
PHOPA	% OF HOUSEHOLDS RECEIVING PUBLIC ASSISTANCE INCOME
PRSSP	% OF HOUSEHOLDS RECEIVING SOCIAL SECURITY
PWCPV	% OF WORKERS COMMUTING BY AUTOMOBILE
PWCPT	% OF WORKERS COMMUTING BY PUBLIC TRANS.
PWALK	% OF WORKERS WALKING TO WORK
PWCOM	% OF WORKERS COMMUTING BY OTHER MEANS
PHWIA	% OF OCCUPIED HOUSING UNITS WITH AT LEAST ONE AUTOMOBILE
P1UHS	% OF ONE UNIT OCCUPIED HOUSING STRUCTURES
P2UHS	% OF TWO UNIT OCCUPIED HOUSING STRUCTURES
P3UHS	% OF THREE OR MORE UNIT OCCUPIED HOUSING STRUCTURES
PHB40	% OF OCCUPIED HOUSING UNITS BUILT BEFORE 1939
P4049	% OF OCCUPIED HOUSING UNITS BUILT BETWEEN 1940 AND 1949
P5059	% OF OCCUPIED HOUSING UNITS BUILT BETWEEN 1950 AND 1959
P6069	% OF OCCUPIED HOUSING UNITS BUILT BETWEEN 1960 AND 1969
P7079	% OF OCCUPIED HOUSING UNITS BUILT BETWEEN 1970 AND 1979
POOHU	% OF OWNER OCCUPIED HOUSING UNITS
PHLIR	% OF OCCUPIED HOUSING UNITS WITH COMPLETE PLUMBING FACILITIES WITH DENSITY LESS THAN 1.0 PERSONS PER ROOM
PHLG1	% OF OCCUPIED HOUSING UNITS WITH COMPLETE PLUMBING FACILITIES WITH DENSITY GREATER THAN 1.0 PERSONS PER ROOM
AGGHV	AGGREGATE VALUE OF OWNER OCCUPIED NON CONDOMINIUM HOUSING UNITS

Sources: Rees, 1968; Berry and Tennant, 1965; Smith, 1973; Liu, 1976; Bederman, 1974; Elgie and Clark, 1980

APPENDIX 3 - DISCUSSION OF PROBLEMS ACCESSING CENSUS TAPES

In order to write a computer program to read census tapes, and to create an output file, it was necessary for me to obtain assistance from a consultant at the University of Tennessee's Computing Center. Even with the consultant's assistance, it still was a difficult process to write a program that would read census tapes. Further, data were unavailable for two of the census tracts. These tracts had to be eliminated from the study. Missing data is not surprising. Tracts with missing data must, however, be removed or the program will not run. It took many changes in the program to find out this problem.

After this problem was corrected, the output was not in a format that could be directly analyzed because many tracts were split. Split tracts were caused by data being reported within and outside of municipal boundaries. I do not understand why totals were not reported for each tract, and the user given the option of obtaining tract totals instead of split tracts. area unit. Because I would be mapping this data, I had to aggregate these split tracts. To do this entailed developing a program written in C to aggregate these split tracts. ¹.

¹. I would like to thank Dr. Ralston for writing a C program to aggregate census tract data.

It is hoped that the the problems that I encountered using census tapes will be eliminated with the 1990 census is available on CD-ROM. I hope that they report totals for each census tract, instead of requiring the user to aggregate split tracts.

In retrospect, I would have saved time if I would have developed the data set manually, but I would have missed out of the experience of learning how to use census tapes.

VITA

Roy Charles Longman was born in Middletown, Ohio on April 12, 1966. He attended school in Middletown, Ohio graduating from high school in June, 1984. In August of 1984, he entered the Middletown campus of Miami University of Ohio. In August of 1986, he transferred to the Oxford campus of Miami University to complete his degree program. He received a Bachelor of Arts Degree in Geography (cum laude) in May, 1988. While at Miami, Mr. Longman was inducted into Phi Beta Kappa, Mortar Board, and Golden Key National Honor Society. He was active in the Student Affairs Council, French Club, Geography Club and as Treasurer of Mortar Board.

Mr. Longman accepted a teaching assistantship and began study at the University of Tennessee, Knoxville, in August, 1988. He received his Masters of Science Degree in Geography in December, 1990.

He is a member of the Association of American Geographers and is currently seeking employment.