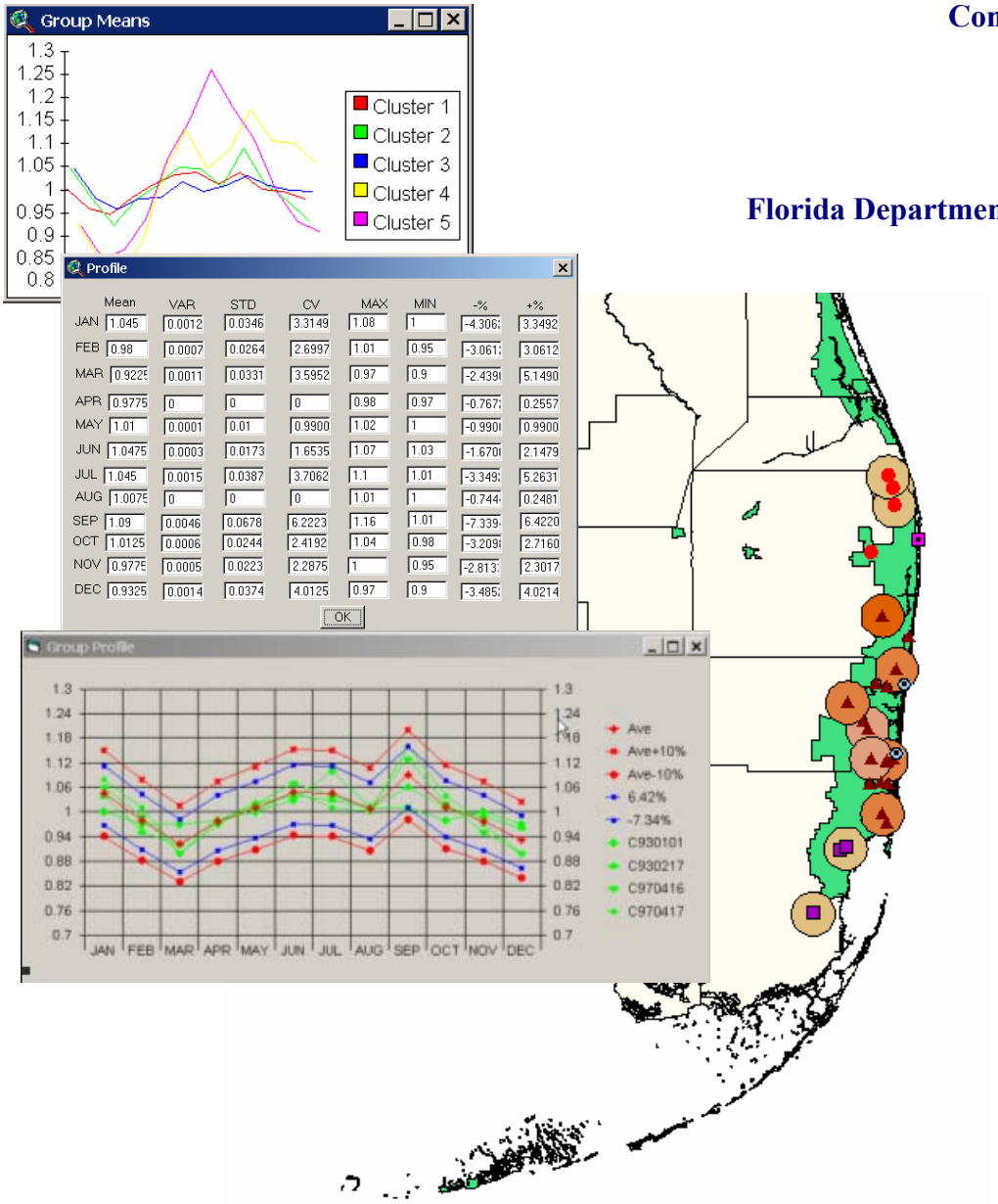
ALTERNATIVES FOR ESTIMATING SEASONAL FACTORS ON RURAL AND URBAN ROADS IN FLORIDA

Final Report
Contract No. BD015-03

Prepared for

Research Office
Florida Department of Transportation



Prepared by

Lehman Center for Transportation Research
Department of Civil & Environmental Engineering
Florida International University

June 2004

Alternatives for Estimating Seasonal Factors on Rural and Urban Roads in Florida

Final Report

Contract No. BD015-03

Prepared for
Research Office
Florida Department of Transportation
605 Suwannee Street, MS 30
Tallahassee, FL 32399-0450

Prepared by

Fang Zhao, Ph.D., P.E.
Associate Professor and Deputy Director

Min-Tang Li, Ph.D.
Senior Research Associate

and

Lee-Fang Chow, Ph.D.
Senior Research Associate

Lehman Center for Transportation Research
Department of Civil & Environmental Engineering
Florida International University
University Park Campus, EAS 3673
Miami, Florida 33199
Phone: 305-348-3821
Fax: 305-348-2802
E-mail: zhaof@fiu.edu

June 2004

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
Final Report for BD015-03		

4. Title and Subtitle	5. Report Date
ALTERNATIVES FOR ESTIMATING SEASONAL FACTORS ON RURAL AND URBAN ROADS IN FLORIDA	June 2004
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Fang Zhao, Min-Tang Li, Lee-Fang Chow	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Lehman Center for Transportation Research, Department of Civil and Environmental Engineering, Florida International University, Miami, Florida 33199	11. Contract or Grant No. BD015-03
12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Research Office Florida Department of Transportation 650 Suwannee Street, MS 30 Tallahassee, Florida 32399-0450	Final Report December 2001 – June 2004
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

In the current practice at the Florida Department of Transportation (FDOT), seasonal factors (SFs) are used in the calculation of annual average daily traffic (AADT) at portable traffic monitoring sites (PTMS). The permanent traffic monitoring sites (TTMSs) are first manually classified into different groups (known as seasonal categories) based on similarities in traffic characteristics of roads and on engineering judgment. FDOT districts then assign a seasonal factor category to each PTMS according to the site's geographical locations, assuming that seasonal variability and traffic characteristics at the short-term and permanent count sites are similar in the same geographic area. The goal of this study was to develop a more objective and data-oriented approach by seeking to explain the underlying causes of seasonal fluctuation patterns in traffic data and to develop a methodology rooted in statistics for assigning seasonal factors to short period counts. Various clustering methods from the nonparametric hierarchical cluster analysis and parametric model-based analysis were evaluated in this study and those that better described the truth-in-data in the seasonal factor grouping process were identified. Conventional multiple linear regression analyses were conducted to identify possible factors contributing to the seasonal fluctuations in traffic volumes. Based on these factors, a fuzzy decision tree was developed to determine the seasonal factor group assignment for short period counts in the southeast Florida urban area. The methodologies developed in this study for seasonal factor grouping and assignment were implemented in a prototype GIS program, which also supported visualization of data on transportation system, traffic data, land use, and statistics about seasonal factors and groupings.

17. Key Word	18. Distribution Statement
Seasonal Factors, AADT, Cluster Analysis, Traffic Monitoring, Regression Analysis, Fuzzy Logic, Decision Tree, Seasonal Factor Group Assignment	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified.	Unclassified.	144	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

DISCLAIMER

The contents of this report reflect the views of the authors and do not necessarily reflect the official views or policies of the Florida Department of Transportation. This report does not constitute a standard, specification, or regulation.

ACKNOWELDGEMENTS

This study was sponsored by a research grant from the FDOT. The authors would like to thank the Project Manager, Mr. Doug O'Hara of the FDOT District 4 of Planning and Environmental Management Office, for his guidance and support throughout the research. The authors also thank Mr. Harshad Desai and Mr. Richard Reel, of the Traffic Data Section of FDOT Transportation Statistics Office, for their valuable input and guidance. The authors are grateful for the useful comments from the Mr. Alexander Rodriguez and Mr. Brent Smith from Planning and Environmental Management Office, FDOT District 4.

Ms. Yifei Wu and Ms. Claudia Lamus, both Graduate Research Assistants of the Lehman Center for Transportation Research during this project, provided assistance in developing the statistical models. Ms. Chris Fraley of the University of Washington answered questions related to the MCLUST software program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS AND ABBREVIATIONS

AADT	Annual Average Daily Traffic
ADT	Average Daily Traffic
AASHTO	American Association of State Highway and Transportation Officials
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
ATR	Automatic Traffic Recorder
AVE	Average linkage method available in SAS
BIC	Bayesian Information Criterion
BP	Back-Propagation
CEN	Centroid method available in SAS
CV	Coefficient of Variation
DelDOT	Delaware Department of Transportation
EM	Expectation-Maximization
EML	Clustering method in SAS assuming spherical multivariate Normal distribution
FDOT	Florida Department of Transportation
FLE	Flexible-Beta Method available in SAS
FTI	Florida Traffic Information
GA	Genetic Algorithm
GDR	Generalized Delta Rule
GIS	Geographic Information System
IDOT	Iowa Department of Transportation
IID	Independently and Identically Distributed
LTPP	long-term pavement performance
MADT	Monthly Average Daily Traffic
MCQ	McQuitty's similarity analysis available in SAS
MED	Median method available in SAS
MLE	Maximum-Likelihood Estimation
MnDOT	Minnesota Department of Transportation
MOCF	Model Output Conversion Factor
MS	Microsoft
MSF	Monthly Seasonal Factor
NCDOT	North Carolina Department of Transportation
PennDOT	Pennsylvania Department of Transportation
PCS	Portable Count Site
PSCF	Peak Season Conversion Factor
PSF	Pseudo *F* statistic
PSWADT	Peak Season Weekday Average Daily Traffic
PTC	Permanent Traffic Counter
PTMS	Portable Traffic Monitoring Site
RCI	Roadway Characteristics Inventory
SAE	Sum of Absolute Error
SAS	Statistical Analysis System
SIN	Single linkage method available in SAS
SF	Seasonal Factor

SHS	State Highway Systems
TMG	Traffic Monitoring Guide
TTMS	Telemetry Traffic Monitoring Site
VIF	Variance Inflation Factor
VDOT	Virginia Department of Transportation
WAR	Ward's minimum-variance method available in SAS

EXECUTIVE SUMMARY

INTRODUCTION

The Florida Department of Transportation (FDOT) stores and reports traffic volume data collected from over 300 telemetry traffic monitoring sites (TTMS) where traffic data are continuously collected and true average annual daily traffic (AADT) information may be obtained. The information collected over these strategically selected locations is utilized to convert short-term traffic counts (also known as coverage counts) collected at portable traffic monitoring sites (PTMSs) where no permanent traffic counters (PTCs) are installed. In the current practice, TTMSs are manually classified into different groups (known as seasonal factor categories) based on similarities in traffic characteristics of roads and engineering judgment. FDOT districts then assign a seasonal factor category to each short-term traffic count site, also known as portable count sites (PCSs), according to the site's geographical location, assuming that seasonal variability and traffic characteristics at the short-term and permanent count sites are similar in the same geographic area. The final seasonal factor groups are often the product of a combination of statistical analysis and analyst's knowledge and expertise.

PROBLEM STATEMENT

The focus of seasonal factor grouping, in short, is to assign labels representing the clusters to the TTMSs, and cluster analysis has been recommended by the Federal Highway Administration (FHWA) to identify roadway sections with similar traffic patterns [TMG01]. Seasonal groups are commonly constructed based on Monthly Seasonal Factors (MSFs). Often the number of seasonal factor groups is unknown, and the same is true as to the group that a TTMS belongs. Conventional nonparametric clustering models such as Ward's Minimum-Variance method have been widely applied in the seasonal factor grouping. However, these methods do not provide guidelines with regards to the choice of the optimal number of groups. Even more problematic is a lack of specifications of definable characteristics to allow the objective assignment of short counts to the seasonal factor groups. The traffic patterns are usually associated with roadways' functions, land use patterns, etc., which may be the dominating factors in determining roadway groupings. Such factors are not well identified or quantified in the current practice. An objective method for seasonal factor grouping and assignment needs to consider these factors to allow the seasonal groups better reflect the unique characteristics of a site as well as to assign the short counts to established seasonal groups in a more rational manner.

NONPARAMETRIC CLUSTERING METHODS

A total of eight agglomerative clustering methods were evaluated in this study for grouping TTMSs. The average linkage, centroid, and single linkage methods were found to be more robust to outliers than the other methods. The study also found that the McQuitty's (MCQ) method performed better than the other methods when grouping TTMSs after outliers were eliminated. Although the results from analyzing the four-year MSF data with the MCQ method showed that the compositions of seasonal groups were not stable over time, the change in the spatially clustering pattern indicated that more variables should be included in the process of determining seasonal cluster groups. The study also led to the finding that roadway function

class did not seem to play an important role in determining seasonal groups in urban areas. Instead, it was the spatial location of a given TTMS that mattered since TTMSs tended to be clustered with those in its proximity.

PARAMETRIC MODEL-BASED CLUSTERING METHODS

The model-based clustering was accomplished using the MCLUST software, an extension of the SAS software. A total of ten models were investigated: EII, VII, EEI, VEI, EVI, VVI, EEE, VVV, EEV, and VEV models. Evaluation of the performance of model-based clustering methods for seasonal factor grouping showed that, without additional information such as the spatial locations of the TTMSs, the model-based clustering methods, such as the EEV model, produced classifications with a negligible grouping error of 2.08% when statewide MSF data were used in the analysis. However, when the geographic locations of the TTMSs were not considered, the resulted clusters included TTMSs located more than 400 miles apart. It is unlikely in practice that TTMSs located so far apart would be grouped together. Therefore, merely using the MSF data was not sufficient to determine the seasonal factor groups when the model-based approach was applied.

By incorporating coordinates of the TTMSs into the model-based clustering, this study found that the EII and VII models produced relatively practical numbers of factor groups. By further comparing the MSFs in the factor groups derived from these two models, the EII model was identified as the one with the best performance since it produced the least grouping error. The results from a systematic analysis of the model-based clustering may be considered as a reasonable starting point for determining the seasonal factor groups in practice. The procedure offers greater flexibility in classifying a TTMS since the probability for a TTMS belonging to a given factor group is estimated. For example, if a TTMS is considered misclassified, it may be easily reassigned to the next factor group according to the sequence determined by the grouping probability. Additionally, incorporating the coordinates of TTMSs in the model-based clustering analysis allows geographical effects to be considered in the grouping process. The groups of TTMSs so derived are not merely similar in their MSF fluctuation patterns but are also spatially clustered together. The results will benefit transportation professionals when assigning a seasonal factor group (category) to a short count site by considering spatial proximity. The model-based clustering process presented in this study may also allow other characteristics such as land use that could not be considered in the conventional grouping approaches to be incorporated in the grouping process.

REGRESSION ANALYSIS FOR URBAN ROADS

Using the MSFs collected from the TTMS sites in Broward, Miami-Dade, and Palm Beach counties and demographic and socioeconomic data mainly from the census, this research identified several significant factors that appeared to contribute to the seasonal patterns of traffic. These factors included concentrations of seasonal residents, tourists (the latter through a variable that reflected concentration of hotels and motels), retired population between age 65 and 75 with high income, and retail employment. Roadway federal functional classification was not found to be a factor. Similarly, no correlation was found between the seasonal factors and traffic volume per lane and number of lanes.

REGRESSION ANALYSIS FOR RURAL ROADS

The MSFs collected from the TTMSs located in FDOT District 2 and 3 were analyzed. It was found that simple buffer methods with various buffer sizes did not capture the underlying causes behind traffic fluctuations over time on rural roads as well as in the case of urban roads. Other variables will need to be identified and incorporated to better quantify the land use as well as socioeconomic/demographic characteristics of the roadway traffic.

CONCLUSIONS AND RECOMMENDATIONS

Seasonal factors are a complex subject. While there have been relatively more studies on various methods to determine seasonal groups, determining the underlying causes of season variations in traffic and developing models to predict seasonal groups has proven to be a significant challenge. So far based on literature, success in explaining or modeling seasonal factors has been limited. This research made contributions to the understanding of the subject by identifying plausible predictors for seasonal groups, further confirming the importance of geographic location in seasonal grouping, providing a theoretical basis for consideration of geographic locations in seasonal factor grouping and assignment, and developing a practical approach for assigning short counts to seasonal groups.

This study first investigated conventional nonparametric hierarchical clustering analysis and parametric model-based cluster analysis methods for seasonal factor grouping. It was found that spatial proximity should be appropriately considered in both grouping and assignment processes. The model-based clustering analyses provided a good starting point for transportation professionals to group TTMSs more accurately into seasonal factor categories in a systematic and data-driven manner by simultaneously considering a TTMS's spatial proximity and their MSFs.

Multiple linear regression analyses were subsequently conducted for selected urban and rural areas to identify possible explanatory variables for seasonal traffic fluctuations. Seasonal residents, tourists, retired people between age 65 and 75 with high income, and retail employment were identified as the significant indicators for seasonal traffic fluctuations on urban roads in southeast Florida. For the rural roads, variables such as functional classification for highways, percentage of seasonal households, agricultural employment, and truck factor were identified as potential explanatory variables.

To develop a methodology to assign a seasonal factor category to a PTMS, a fuzzy decision tree was constructed using the TTMS groups obtained from the model-based cluster analysis and based on the aforementioned four variables for the tri-county urban area, i.e., Broward, Miami-Dade, and Palm Beach counties. The decision tree was then applied to determine the seasonal factor category for a given PTMS. The decision tree was easy to visualize and apply, and the assignment results were self-explanatory. For example, areas with a larger number of visitors and a larger number of seasonal households would expect to experience more fluctuation in traffic volumes.

A GIS based computer program was developed as part of this research to demonstrate the usefulness of a GIS user interface for visualization of land use, demographic, and socioeconomic data, as well as the characteristics of the transportation systems and traffic counts. Buffer analysis, regression analysis, and cluster analysis were also supported in the program for advanced users who are interested in performing statistical analysis. The statistical functions were provided by SAS and S-Plus.

Although this study developed regression models that could potentially be used to estimate seasonal factors directly for a PTMS, because of the limited sample size, the predictive power of the models could not be determined. Additionally, because traffic in different urban areas may have different seasonal patterns due to differences in climate, local economy, and demographics, variables identified in this study may not be directly applicable to other areas.

The following recommendations were made based on the findings from this research:

- To make the results from this research useful to all FDOT districts, where the seasonal categories are determined and assigned to PTMSs, and even to local government users who operate a local traffic statistics program, additional studies need to be carried out to determine whether the variables identified in this study for the urban areas in southeast Florida are also applicable to other urban areas in the state. Due to differences in local land use patterns and economies, it is possible that some urban areas have a different set of variables that explain the patterns of traffic variations.

- The regression models for estimating MSFs for rural roads currently have relatively low R^2s. To improve the model performance and identify better MSF predictors, further analyses are necessary. They may include the development and testing of improved or new variables and new modeling techniques such as nonlinear regression models.

- A standard procedure should be developed by FDOT based on the results from this study and future studies. This standard procedure should be based on a set of statistics based methods for seasonal factor grouping and assignment that are more objective and data-driven and that minimize the reliance on individuals' experience and subjective judgment. Such a standard procedure will help improve the quality of the transportation data used in important decision making processes.

- The current prototype GIS program is a demonstration program developed for FDOT District 4. It needs to be expanded to include all FDOT districts. The program and the necessary data need to be delivered in a single CD-ROM, similar to the current traffic CD-ROM published by FDOT each year. The data required by the program, which are from the U.S. Census Bureau and from urban area travel demand models, need to be made available from the Internet. A possible central depository location may be the Florida Geographic Digital Library (FGDL) at the University of Florida. The data should be updated every three to five years as more recent data become available or when census data are released.

- The current GIS program is implemented in the ArcView environment. When the FDOT district offices and central office completely migrate to ArcGIS, this program may be re-implemented by customizing ArcGIS with VBA (the programming language in ArcGIS). Alternatively, a program implemented in MapObject (also an ESRI product) may be developed. The advantage of a MapObject based program is that it does not require any GIS software from the user and still provides the same GIS functionalities. A MapObject based program will allow the GIS program to be distributed to the users on a CD and used in the same way as the Traffic Data CD.

1. INTRODUCTION

One of the major responsibilities of state departments of transportation is to collect and store traffic data. These data are used as inputs to numerous types of analyses, including roadway design, pavement design, air quality, and maintenance. In the past, different states applied different approaches to analyze the data that the agencies collected from the field. In a 1991 article, Albright raised issues related to the varied practices in different states on traffic data collection and statistics reporting [ALB91]. He presented an imperative need for national traffic monitoring standards and guidelines. In his 1993 article, Albright addressed the issues of pattern identification to validate traffic data from permanently installed and continuously operating traffic recoding devices [ALB93]. Subsequently, guidelines on traffic data collection and reporting were standardized and more permanent traffic monitoring devices were installed on roadways to monitor traffic.

It is well known that traffic variations occur at different time scales, e.g., time of day, day of week, and season (such as month) of the year, as stated in the Traffic Monitoring Guide [TMG01], a report published by the Federal Highway Administration (FHWA). Of the known temporal fluctuations of traffic stream, seasonal variation is probably the most important characteristic that must be accounted for in traffic monitoring. Currently, the Florida Department of Transportation (FDOT) stores and reports traffic volume data collected from over 300 telemetry traffic monitoring sites (TTMS), from which the true average annual daily traffic (AADT) information may be obtained. The information collected over these strategically selected locations is utilized to convert short-term traffic counts (also known as coverage counts) collected at portable traffic monitoring sites (PTMSs) where no TTMSs are installed.

In Florida, four factors are used in converting short-term traffic counts to traffic volumes for different purposes. They are weekly Seasonal Factors (SF), Peak Season Conversion Factors (PSCF), Model Output Conversion Factors (MOCF), and Axle Correction Factors. PSCF is used to convert a short-term traffic count (ADT) to peak season weekday average daily traffic (PSWADT) and MOCF is used to convert the PSWADT to average annual daily traffic (AADT). Among these factors, SF plays a key role in estimating traffic volumes since this factor is used in calculating not only Average Annual Daily Traffic (AADT) but also PSCF and MOCF. It is crucial to properly consider and accurately interpret the temporal variation effects on collected traffic data in order to achieve better design decisions.

1.1 Background

In the current practice, the FDOT first calculates the SFs on Florida's roadway segments for each week of the year at each TTMS. The SFs are determined by interpolating between the monthly seasonal factors (MSFs) for two consecutive months. The MSF for a specific month at a particular location is derived from dividing the monthly average daily traffic (MADT) at a given location with its AADT. The TTMSs are then manually classified into different groups (known as categories) based on similarities in traffic characteristics of roads and engineering judgment. For example, there are 178 SF categories based on the 1999 traffic data. The weekly SFs for a specific category are subsequently obtained by calculating the arithmetic averages of the factors from the TTMSs in the same group during the same period of time. FDOT districts then assign a

seasonal factor category to each short-term traffic count site, also known as portable count sites (PCSs), according to the site's geographical location, assuming that seasonal variability and traffic characteristics at the short-term and permanent count sites are similar in the same geographic area. The final factor group definition is often a combination of statistical analysis and analyst knowledge and expertise.

1.2 Problem Statement

There are two important components consisted in the factoring process for seasonal factors: the determination of the TTMS categories and the assignment of SF categories to PCSs. Currently in Florida, TTMSs are grouped according to subjective criteria, and only the geographic location of a short-term count site and its functional classification is considered when a SF category is assigned.

The problem of constructing factor groups from TTMSs and estimating monthly factors with a given precision has attracted a lot of attention over the years. Numerous studies have been conducted in the past to identify alternative approaches to reveal the truth-in-data and to reduce subjective judgment involved in traffic data analysis. The focus of seasonal factor grouping, in short, is to assign labels representing the clusters to the TTMSs, and cluster analysis has been recommended by the FHWA for identifying roadway sections with similar traffic patterns [TMG01]. Seasonal groups are commonly constructed based on MSFs. Often the number of seasonal factor groups is unknown, as is to which cluster that a TTMS belongs. Conventional nonparametric clustering models such as Ward's Minimum-Variance method have been widely applied in the seasonal factor grouping. However, these methods lack theoretical guidelines on establishing the optimal number of groups.

The major difficulty in developing factors groups lies not in the aggregation of the continuous counters to a given group, but rather in the specification of definable characteristics to allow the objective assignment of short counts to the seasonal factor groups. The traffic patterns, however, are usually associated with roadways' functional classifications (such as rural, urban, interstate, collector, and recreational), land uses, etc., which may be important factors in determining roadway groupings. These factors are not well quantified in the current practice in Florida. By appropriately considering and incorporating these factors into the data collection and processing, it is possible to reduce the data collection effort while improving the accuracy of SF estimations.

1.3 Goal and Objectives

The goal of this research is to incorporate new technologies to enhance the current factoring process in traffic monitoring for estimate traffic volumes on Florida's urban and rural roads. The objectives of this research are described as follows:

1. Identify, evaluate, and develop alternative approaches that have the potential of improving the current seasonal factor grouping process.

2. Identify possible explanatory variables that allow more accurate assignment of short-count sites to a given seasonal factor group.

3. Develop a methodology to assign established seasonal factor groups to short count sites based on the explanatory variables.

1.4 Organization

This document first summarizes various methods for incorporating seasonal variations in the calculation of AADT in Chapter 2, including conventional approaches such as statistical cluster analysis, geographic/functional assignment, and regression analysis, as well as machine learning techniques such as neural networks and genetic algorithms. The existing literature on assigning a short count site to a seasonal group is also reviewed. The traffic data that are used in the analysis are then briefly described in Chapter 3. In Chapter 4, the performances of nonparametric hierarchical cluster methods and parametric model-based cluster methods for classifying TTMSs into seasonal factor groups are assessed. Chapter 5 describes the land use characteristics identified via multiple linear regression analysis and geographical locations of a count stations, which may help in determining the seasonal factor category for a short-count site. Finally, conclusions and recommendations are provided in Chapter 6.

In this study, unless explicitly stated otherwise, permanent ATR stations, telemetry traffic monitoring sites (TTMS), and permanent traffic count (PTC) sites all refer to permanent traffic monitoring devices used for continuously recording traffic flows and providing day-to-day traffic information throughout a year. Portable traffic monitor sites (PTMSs), short-count stations, and portable count stations (PCSs) all refer to the count stations temporarily installed on roadways to collect coverage counts.

2. LITERATURE REVIW

This chapter provides a summary of the research efforts in the past to enhance the factoring process to improve the accuracy in the estimation of traffic volume at a short count station. As mentioned in the previous chapter, the factoring process consists of both grouping and assigning procedures. Up to present, most of the research efforts focused on seasonal factor grouping and numerous approaches were proposed to obtain better groupings in the factoring process. These approaches differ mainly in their data and processes. The 2001 Traffic Monitoring Guide (TMG) suggests the following three alternative techniques for determining factor groups [TMG01]:

- **Cluster analysis**. A least squares-minimum distance algorithm is used to determine the variation patterns in the data from TTMSs. The count sites that are determined to be most similar are grouped together, and the process is repeated to determine the next similar group. According to TMG, this is the best approach to determine the grouping of permanent traffic counters since it avoids subjective factors involved in the analysis and is based on sound statistical procedures.

- **Geographic/functional assignment of roads to groups**. Roads are classified into factor groups on the basis of a combination of geographic location and functional roadway classification.

- **Same road application of factors**. The factor from a single PTC is assigned to all road sections within the influence zone of that count site. The influence zone is a road area in which the characteristics of traffic volumes do not change significantly.

In addition to the above three techniques, models developed based on conventional linear regression, neural networks, and genetic algorithms are also reported in the literature for grouping permanent traffic count sites. In the following sections, the current practice in Florida of estimating the seasonal factor at a given short count station is first described, following by a discussion of numerous modeling techniques relevant to the factoring process including grouping and assignment procedures.

2.1 Current Practice in Florida

Currently, the FDOT applies the following equation to estimate the AADT at each TTMS [PTFH02]:

$$AADT = ADT \times SF \times Axle \quad (1)$$

where

Axle = axle correlation factor that converts the counted number of axels to the number of vehicles;
ADT = average daily traffic, typically the average value of a 72-hour traffic count collected from Tuesday to Thursday;
SF = seasonal factor that reflects traffic seasonal fluctuation pattern; and

AADT = estimate of typical daily traffic on a road segment for all days of the week, Sunday though Saturday, over the period of one year.

In the current practice, a small number of TTMSs are manually grouped into clusters or factor groups according to the similarities in their monthly variation patterns. Each coverage count site is then assigned one of these factor groups. The associated AADT for a given coverage count location is then estimated with the seasonal factor computed for the assigned factor group. By multiplying seasonal factor (SF) and axle correction factor with ADT, the estimated AADT is expected to be statistically accurate if the SF is accurate. For example, based on the data from an ATR station maintained by the Iowa DOT (IDOT) in Cedar Rapids, Iowa, it was shown that a 25% error reduction was achieved by incorporating day-of-week and month-of-year traffic variations into the AADT prediction from a short-term traffic count [GRA98]. The same or similar procedure is applied for estimating AADT at short count sties nationwide. The drawback of such a factoring process is that subjective criteria are applied in the manual grouping and assigning procedures and the seasonal factors may not reveal the truth-in-data from the collected traffic data.

2.2 Cluster Analysis

The purpose of cluster analysis is to place TTMSs into groups such that TTMSs in a given cluster have similar seasonal fluctuations. Cluster analysis is a technique that does not make assumptions about the number of groups or the group structure [JOH02]. Grouping is achieved on the basis of similarities measured as distance, i.e., dissimilarities. As such, input to cluster analysis is usually data from which similarities may be measured. In the context of seasonal factor grouping, input to cluster analysis is usually 12 monthly seasonal factors (MSFs) for each TTMS.

There are several types of statistical cluster methods for grouping objects. Among these methods, nonparametric methods, including agglomerative hierarchical clustering and nonhierarchical clustering, have been typically used in determining seasonal factor groups in the practice. The parametric model-based clustering approach, however, has now become popular in a variety of disciplines in determining cluster membership. The following sections describe the applications.

2.2.1 Nonparametric Agglomerative Hierarchical Clustering Methods

Nonparametric clustering classifies objects into categories based on a measure of similarity between clusters. The basis of nonparametric clustering is that groups correspond to modes of an unknown distribution function. Consequently, the goal is to estimate the modes and assign each observation to the domain of attraction of a mode. The nonparametric agglomerative hierarchical cluster analysis process (refer to as the hierarchical cluster analysis hereafter) begins by treating each observation as a cluster by itself. The two closest clusters determined by a specific similarity measure are merged to form a new cluster to replace the two old clusters. Merging of the two closest clusters is repeated until only a single cluster remains. The nonparametric agglomerative hierarchical cluster analysis methods organize objects so that one

cluster may be entirely contained within another cluster and no other kind of overlap between clusters is allowed.

In cluster analysis, similarity or closeness between two *p*-dimensional observations is usually measured by Euclidean distance:

$$d(\mathrm{x},\mathrm{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_p - y_p)^2} = \sqrt{(\mathrm{x}-\mathrm{y})^{\mathrm{T}}(\mathrm{x}-\mathrm{y})} \tag{2}$$

where $\mathrm{x} = [x_1, x_2, \ldots, x_p]^{\mathrm{T}}$ and $\mathrm{y} = [y_1, y_2, \ldots, y_p]^{\mathrm{T}}$. Other distance measures, including Minkowski metric, Canberra metric, and Czekanowski coefficient, are defined as follows [JOH02]:

Minkowski metric:
$$d(\mathrm{x},\mathrm{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m\right]^{\frac{1}{m}} \tag{3}$$

Canberra metric:
$$d(\mathrm{x},\mathrm{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{x_i + y_i} \qquad (x_i, y_i \geq 0) \tag{4}$$

Czekanowski coefficient:
$$d(\mathrm{x},\mathrm{y}) = 1 - \frac{2\sum_{i=1}^{p} \min(x_i, y_i)}{\sum_{i=1}^{p} (x_i + y_i)} \qquad (x_i, y_i \geq 0) \tag{5}$$

The Minkowski metric provides the same similarity measure as the Euclidean distance when *m* is equal to 2, and by using different values of *m*, the weights given to the distances may be manipulated. For the Canberra metric and Czekanowski coefficients, only nonnegative variables may be defined. Currently, most commercially available software packages for clustering analysis only utilize the Euclidean distances as the similarity measure.

Nonparametric hierarchical clustering analysis is used to determine the groupings in the data. In the case of seasonal factoring process, the seasonality observed at each TTMS from month to month is considered in the grouping process. The basic intent of the cluster analysis is to identify variation patterns to give the analyst the knowledge and insight to develop grouping criteria to expand short counts to AADT.

Various clustering methods have been employed in seasonal factor analysis. For example, Sharma and Werner applied a hierarchical clustering method to group 45 PTCs in Alberta, Canada based on their 12 monthly factors [SHA81, SHA83]. The Scheffe's S-method of multiple comparisons of group means was used to determine the optimal number of groups ranging from 6 to 10 obtained from the hierarchical process, each containing more than two counters. The results showed that eight to nine groups were desirable. Sharma and Allipuram then applied the method proposed in the previous work [SHA81] to group 61 PTCs in Alberta again using the data collected in 1989 and obtained a total of seven cluster groups [SHA93].

Sharma *et al*. concluded in their later work that the AADT estimation errors were more sensitive to the correctness of sample site assignment to a proper PTC group [SHA96].

Aunet used cluster analysis to examine the variation in Wisconsin's traffic data [AUN00]. The procedure consists of the following steps:

1. Examining plots of monthly traffic at each permanent count station;
2. Examining tables of coefficient of variations (CVs) for each permanent count station;
3. Examining the results of cluster analysis; and
4. Examining geographic mapping of ATRs in preliminary groupings.

The preliminary results from Aunet's procedure applied in Wisconsin revealed that seasonal patterns remained stable over time. Additionally, although significant variations existed in the monthly seasonal factors for the permanent count stations classified into the same group, seasonal factor groups could be generally defined according to roadway functional classifications, i.e., urban, rural, and recreational.

There are numerous hierarchical clustering methods available in Statistical Analysis System (SAS) to determine which individual elements or clusters should be merged together [SAS99]. The various clustering methods differ in how the Euclidean distance between two clusters is computed. The SAS CLUSTER procedure provides the following algorithms for agglomerative hierarchical grouping [SAS99]:

1. Average Linkage (AVE)
2. Centroid Method (CEN)
3. Maximum-likelihood for mixtures of spherical multivariate normal distributions with equal variances but possibly unequal mixing proportions (EML)
4. Flexible-beta Method (FLE)
5. McQuitty's Similarity Analysis (MCQ)
6. Median Method (MED)
7. Single Linkage (SIN)
8. Ward's Minimum-Variance Method (WAR)

Table 1 shows the distance definitions for these hierarchical clustering methods. In SAS, options for squared and non-squared Euclidean distances may be specified.

The process of cluster analysis is completely driven by the variability in the MSFs. Two apparent advantages of cluster analysis are that it allows for independent determination of "similarity" between groups, thus making the groups less subject to bias, and that it is able to identify travel patterns that may not be intuitively obvious to the analyst [TMG01]. Thus, it helps agency staff investigate road groupings they might not otherwise examine, which in turn may lead to more efficient and accurate factor groups and providing new insights into travel patterns.

Table 1. Standard Hierarchical Clustering Methods

Method	Distance between Clusters
AVE	Average distance between pair of objects, each in a different cluster
CEN	Distance between centroids in two clusters
EML	Maximum-likelihood hierarchical clustering for mixtures of spherical multivariate normal distributions with equal variances
FLE	A combined method: specification of a value of -1.0 results in complete linkage, a value of 1.0 yields single linkage with extreme chaining, and a value near -0.25 approximates average linkage
MCQ	Distance between clusters is weighted using arithmetic averages
MED	Squared Euclidean distance between weighted centroids
SIN	Minimum distance between pair of objects, one in one cluster, one in the other
WAR	Increase in sum of squares within clusters

Hierarchical clustering analysis also has its shortcomings in that it provides no definable characteristic or criteria upon which to form groups. Consequently, although well adopted in the practice, this type of clustering applications suffers from the following two major weaknesses [TMG01]:

- *Lack of theoretical guidelines on establishing the optimal number of groups*. It is often difficult to determine how many groups should be formed. The difficult task is to determine at what point the sequential merging process should stop. Unfortunately, the "optimal" number of groups cannot be determined mathematically. Consequently, the results of the cluster analysis may not be the ultimate answer. Modifications are to be expected. Statistical models may be used to better understand the variation of data by identifying the seasonal fluctuation patterns and eliminating stations with extreme variations. However, the development of the final factor groups must account for variability and also include characteristics that define the groups to allow the assignment of short counts to the groups in the subsequent process. The knowledge of other criteria, e.g., functional class, geography, topography, degree of urbanization, etc., and the use of analytical judgment are still necessary in interpreting the results.

- *Lack of theoretical guidelines on group assignment*. The formed groups often cannot be adequately defined, because the cluster procedure considers only the traffic variability at TTMSs, which cannot be directly applicable to the short counts. Plotting on a map for the sites that fall within a specific cluster group is sometimes helpful when attempting to define a given group output by the cluster process. However, in some cases, the purely mathematical nature of the cluster process simply does not lend itself to easily identifiable groups. No criteria for assignment of short counts to the groups have been defined via the hierarchical cluster analysis. This is where the descriptive analysis and the use of functional class, geography, or topography are needed to provide additional criteria for assignment formation.

2.2.2 Nonhierarchical Clustering Methods

Nonhierarchical (also known as partitioning) clustering methods place each object in only one cluster. The methods usually begin by randomly partitioning individual items into *k* groups to avoid any overt biases. Items are then assigned to clusters with the nearest medium or mean. The number of clusters (*k*) may be either given as a priori or determined by the algorithm. When *k* is unknown, nonhierarchical methods are generally repeated for several values of *k*. The optimal value evaluated by the criterion associated with each nonhierarchical method is then selected as the desired number of groups. Since the *k* clusters are generated simultaneously, the resulted classification is non-hierarchical. However, a hierarchy of nonhierarchical classification may be constructed using the results repeatedly with several values of *k* [MAS89]. For this reason, the definition of nonhierarchical clustering is vague. Nonhierarchical clustering refers to methods that are commonly known as *k*-mean methods. In SAS, the CLUSTER procedure provides the following two models for nonhierarchical clustering:

1. Density linkage, including Wong's hybrid and k^{th}-neareast neighbor methods; and
2. Two-stage density linkage.

Flaherty used the hierarchical clustering method and the *k*-means method available in the Systat software package for microcomputers to analyze the monthly factor data collected over a five-year period from 28 PTCs installed in Arizona [FLA93]. The *k*-means algorithm in Systat was used to produce clusters of prescribed numbers, varying from two to nine, by maximizing the ratio of between-cluster variation to within-cluster variation. This approach was analogous to a one-way ANOVA seeking the largest F-value by reassigning objects.

The results from the hierarchical clustering analysis of Flaherty's study were inconclusive. Flaherty, however, claimed that the results from the nonhierarchical clustering analysis were more straightforward and easier to interpret. Two count stations were found to be consistent outliers over the five-year period and were thus excluded from the analysis. Traffic volumes on Monday, Tuesday, Wednesday, and Thursday were then randomly selected from the remaining 26 PTCs as the surrogates for short-term traffic counts. For comparison purposes, these simulated short counts were adjusted in four different ways to obtain AADT estimates, i.e., AADTs adjusted by the PTCs' own monthly factors and by the appropriate group monthly factors derived from the cluster analysis for three, four, and five clusters. The results of using these simulated factored short counts to estimate AADT were then compared on the basis of standard deviations and coefficients of variation.

Flaherty concluded that similarity in the patterns of the monthly factors was more a function of geography and topography than functional classification of the highways on which the count stations were located and that the population of the surrounding area did not appear to be an explanatory factor for the factor groups. Flaherty also found that four clusters were the best and the most stable of all the variations used in the analysis. Similar to hierarchical clustering methods, difficulties were encountered as how to appropriately interpret the resulted groups from nonhierarchical clustering methods and how to conduct short count site assignments.

2.2.3 Model-Based Gaussian Cluster Analysis

Model-based clustering assumes that each seasonal factor group may be represented by a density function that is a member of some parametric family, e.g., the multivariate normal (Gaussian) family, and that the associated parameters may be estimated from observations [FRA98]. The fundamental concept of model-based clustering analysis is to determine the probabilistic density function for the k^{th} seasonal factor group by estimating the first two orders of statistics, i.e., the p-dimensional mean vector ($\boldsymbol{\mu}_k$) and the $p \times p$ covariance matrix ($\boldsymbol{\Sigma}_k$). If $\boldsymbol{\Sigma}_k$ is expressed in terms of its eigenvalue decomposition, i.e., $\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^{\mathrm{T}}$, where superscript T denotes matrix transpose and $\mathbf{D}_k$, λ_k, and $\mathbf{A}_k$ govern the orientation, the volume, and the shape for the k^{th} seasonal factor group, a systematic analysis may be performed by treating these geometric features as different parameters. Examples of models include $\lambda \mathbf{I}$, $\lambda_k \mathbf{I}$, and $\lambda \mathbf{D}\mathbf{A}\mathbf{D}^{\mathrm{T}}$, etc.

In model-based methods, a maximum-likelihood criterion is used to merge groups. Two approaches are commonly applied in model-based clustering analysis: the classification approach and the mixture approach [DUN02]. The classification approach aims at maximizing the likelihood over the mixture parameters and identifying the group to which each sample belongs. The mixture approach merely aims at maximizing the likelihood over the mixture parameters. Different from a discrete value indicating the cluster in the classification approach, a probability is obtained for a given observation that is classified to a specific group in the mixture approach, and the sum of the probabilities is equal to 1. Compared to non-parametric clustering methods, the ability to estimate the number of groups is an important strength of the model-based approach. Fraley and Raftery employed Bayesian Information Criterion (BIC) with a penalty for the complexity of the model subtracted from the mixture log likelihood to find the optimal number of clusters [FRA98]. The BIC may be used to systematically compare models with different parameterizations, different numbers of seasonal factor groups, or both.

The background of the model-based cluster analysis for seasonal factor grouping is briefly described as follows [TAN02]. Assuming there are G seasonal factor groups in a given study area. For each permanent count station i, the MSF for every month in a year (or a linear combination of these factors) and other characteristics form a p-dimensional vector, $\mathbf{x}_i$. Given $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, where n is the number of PTCs, the density function for the i^{th} PTC from the k^{th} seasonal factor group is $f_k(\mathbf{x}_i \mid \boldsymbol{\theta}_k)$ with some unknown vector of parameters $\boldsymbol{\theta}_k$, where $\boldsymbol{\theta}_k$ consists of a mean vector $\boldsymbol{\mu}_k$ of length p for the mean in each dimension and a $p \times p$ covariance matrix $\boldsymbol{\Sigma}_k$. Assuming $f_k(\mathbf{x}_i \mid \boldsymbol{\theta}_k)$ is multivariate normal (Gaussian), the probability density function has the following form:

$$f_k\left(\mathbf{x}_i \middle| \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) = \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)\right\}}{(2\pi)^{p/2} \left|\boldsymbol{\Sigma}_k\right|^{1/2}} \tag{6}$$

Each seasonal factor group forms an ellipsoid that is centered at its means $\boldsymbol{\mu}_k$ with its geometric characteristics determined by the covariance matrix $\boldsymbol{\Sigma}_k$. The covariance matrix may be expressed in terms of its eigenvalue decomposition as follows [BAN93]:

$$\boldsymbol{\Sigma}_k = \mathbf{D}_k \boldsymbol{\Lambda}_k \mathbf{D}_k^{\mathrm{T}} \tag{7}$$

where

$\mathbf{D}_k$ = orthogonal matrix of eigenvectors, which determines the orientation of $\boldsymbol{\Sigma}_k$; and
$\boldsymbol{\Lambda}_k$ = a diagonal matrix with the eigenvalues of $\boldsymbol{\Sigma}_k$ on the diagonal, which specifies the size and shape of the density contours.

$\boldsymbol{\Lambda}_k$ may be furthered decomposed as follows:

$$\boldsymbol{\Lambda}_k = \lambda_k \mathbf{A}_k \tag{8}$$

where

λ_k = the first eigenvalue of $\boldsymbol{\Sigma}_k$, which specified the volume of the k^{th} seasonal factor group; and
$\mathbf{A}_k$ = $\left[\alpha_{1k}, \cdots, \alpha_{pk}\right]^{\mathrm{T}}$, $1 = \alpha_{1k} \geq \alpha_{2k} \geq \ldots \geq \alpha_{pk} > 0$.

Consequently, Equation (8) becomes:

$$\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^{\mathrm{T}} \tag{9}$$

$\mathbf{D}_k$, λ_k, and $\mathbf{A}_k$ govern the orientation, the volume occupied by the cluster in p-space, and the shape for the k^{th} seasonal factor group, respectively. By treating these geometric features as independent sets of parameters, a systematic analysis may be carried out by constructing models with different parameters. Table 2 shows the models proposed in the context of cluster analysis for covariance matrices [FRA02]. In Table 2, the model identifiers code geometric characteristics of the model. For example, EVI denotes a model in which the volumes of all clusters are equal (E), the shapes of the clusters may vary (V), and the orientation is the identity (I). Clusters in this model have diagonal covariances with orientation parallel to the coordinate axes. Parameters that are associated with characteristics designated by E or V may be determined from the data.

The common heuristic agglomerative clustering algorithms, e.g., average linkage, single linkage, complete linkage, and Ward's method, are each equivalent to a model-based method [KAM02]. More specifically, under the assumption that every $\boldsymbol{\Sigma}_k$ is independently and identically distributed (IID) normal variants, i.e., $\boldsymbol{\Sigma}_k = \lambda I$ (the EII model in Table 1), every seasonal factor group would have the same shape, volume, and orientation since $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma} = \lambda \mathbf{DAD}^{\mathrm{T}}$, which is commonly known as the Ward's method of the conventional clustering approach [KAM02]. The model for the composite of the clusters is usually formulated by the classification likelihood approach or the mixture likelihood approach. The following sections describe the background of these two model-based approaches.

Table 2. Available Parameterizations of Covariance Matrix

Model	Identifier	Distribution	Volume	Shape	Orientation
$\lambda\mathbf{I}$	EII	Spherical	Equal	Equal	NA
$\lambda_k\mathbf{I}$	VII	Spherical	Variable	Equal	NA
$\lambda\mathbf{A}$	EEI	Diagonal	Equal	Equal	Coordinate Axes
$\lambda_k\mathbf{A}$	VEI	Diagonal	Variable	Equal	Coordinate Axes
$\lambda\mathbf{A}_k$	EVI	Diagonal	Equal	Variable	Coordinate Axes
$\lambda_k\mathbf{A}_k$	VVI	Diagonal	Variable	Variable	Coordinate Axes
$\lambda\mathbf{D}\mathbf{A}\mathbf{D}^{\mathrm{T}}$	EEE	Ellipsoidal	Equal	Equal	Equal
$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^{\mathrm{T}}$	VVV	Ellipsoidal	Variable	Variable	Variable
$\lambda\mathbf{D}_k\mathbf{A}\mathbf{D}_k^{\mathrm{T}}$	EEV	Ellipsoidal	Equal	Equal	Variable
$\lambda_k\mathbf{D}_k\mathbf{A}\mathbf{D}_k^{\mathrm{T}}$	VEV	Ellipsoidal	Variable	Equal	Variable

2.2.3.1 *Classification Likelihood Approach*

In the classification likelihood approach, the objective is to identify the parameters $\boldsymbol{\theta}$ and labels $\boldsymbol{\gamma}$ that maximize the following likelihood function:

$$\mathrm{L_C}(\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_\mathrm{G};\gamma_1,\cdots,\gamma_\mathrm{n}\,|\,\mathbf{x}) = \prod_{i=1}^{n} f_{\gamma_i}\left(\mathbf{x}_i\,|\,\boldsymbol{\theta}_{\gamma_i}\right) \tag{10}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^{\mathrm{T}}$ denotes the identifying labels for the classification, i.e., $\gamma_i = k$ for the i^{th} PTC that is classified to the k^{th} seasonal factor group. The presence of the class labels in the classification likelihood introduces a combinatorial aspect that makes exact maximization impractical [FRA02]. Consequently, model-based hierarchical clustering methods are commonly implemented since they usually provide a good approximation of the optimal grouping and are relatively easy to compute [FRA96]. The process is to successively merge a pair of clusters that yields the greatest increase in maximum likelihood expressed in Equation (10). The resulting partitions are suboptimal since the final results may not be global optimal.

2.2.3.2 *Mixture Likelihood Approach*

The objective function in the mixture likelihood clustering approach is to identify the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ that maximize the following likelihood function:

$$\mathrm{L_M}(\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_\mathrm{G};\tau_1,\cdots,\tau_\mathrm{G}\,|\,\mathbf{x}) = \prod_{i=1}^{n}\sum_{k=1}^{G}\tau_k\, f_k\left(\mathbf{x}_i\,|\,\boldsymbol{\theta}_k\right) \tag{11}$$

where $\boldsymbol{\tau}_k$ is the probability that a PTC belongs to the k^{th} seasonal factor group that meets the following constraints:

$$\tau_k \geq 0 \tag{12}$$

$$\sum_{k=1}^{G} \tau_k = 1 \tag{13}$$

In the mixture likelihood approach, it is assumed that there exists a finite set of G seasonal factor groups and each PTC is associated with an indicator vector $\mathbf{z}_i$ of length G whose components are all zero except for one indicating the classification. The key difference between the classification and mixture approaches is that in the former each PTC is assigned to a unique cluster, while in the latter each PTC is assigned with a probability of originating from each seasonal factor group. Moreover, the mixture approach allows the uncertainties associated with the class membership of the observations to be estimated. The equivalent log-likelihood function of Equation (11) is:

$$l_{\mathrm{M}}(\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_{\mathrm{G}};\tau_1,\cdots,\tau_{\mathrm{G}}\,|\,\mathbf{x}) = \sum_{i=1}^{n} \ln\left(\sum_{k=1}^{G} \tau_k\, f_k\left(\mathbf{x}_i\,|\,\boldsymbol{\theta}_k\right)\right) \tag{14}$$

Equation (14) may be optimized over τ_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ using the expectation-maximization (EM) algorithm. The EM algorithm is a general approach to maximum-likelihood estimation (MLE) in the presence of incomplete data. The complete data are considered to be $\mathbf{y}_i = (\mathbf{x}_i, \mathbf{z}_i)$, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})$ constitutes the "missing" data and z_{ik} is equal to one for a PTC ($\mathbf{x}_i$) belonging to seasonal factor group k and zero otherwise. Equation (14) is thus considered as the log likelihood function from the observed data $\mathbf{x}_i$. Assuming that each $\mathbf{z}_i$ is independent and identically distributed according to a multinomial distribution of one draw from G seasonal factor groups with unknown probabilities $\tau_1, \ldots, \tau_G$, the probability mass function for the i^{th} PTC (i.e., $\mathbf{x}_i$) belonging to seasonal factor group k may be expressed as follows [DUN02]:

$$f(\mathbf{z}_i) = \frac{1!}{0!\cdots 1!\cdots 0!}\tau_1^0\cdots\tau_{k-1}^0\tau_k^1\tau_{k+1}^0\cdots\tau_G^0 = \tau_k \tag{15}$$

Assuming the probability density function for $\mathbf{x}_i|\mathbf{z}_i$ (i.e., $\mathbf{x}_i$ given $\mathbf{z}_i$) as

$$f\left(\mathbf{x}_i\,|\,\mathbf{z}_i\right) = \prod_{k=1}^{G} f_k\left(\mathbf{x}_i\,|\,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)^{z_{ik}} \tag{16}$$

Combining Equations (15) and (16) to obtain the probability density function for $\mathbf{y}_i$ yields

$$f\left(\mathbf{y}_i\right) = f\left(\mathbf{x}_i\,|\,\mathbf{z}_i\right)\times f(\mathbf{z}_i) = \prod_{k=1}^{G} f_k\left(\mathbf{x}_i\,|\,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)^{z_{ik}}\tau_k \tag{17}$$

Under the condition that z_{ik} is equal to one for $\mathbf{x}_i$ belonging to seasonal factor group k and zero otherwise, Equation (17) may be generalized as follows:

$$f\left(\mathbf{y}_i\right) = \prod_{k=1}^{G}\left(f_k\left(\mathbf{x}_i\,|\,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)\tau_k\right)^{z_{ik}} \tag{18}$$

For a total of *n* PTCs, Equation (18) may be written as

$$f(\mathbf{y}) = \prod_{i=1}^{n} \prod_{k=1}^{G} \left(f_k\left(\mathbf{x}_i \middle| \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tau_k \right)^{z_{ik}} \tag{19}$$

The resulted complete-data log likelihood is

$$l(\boldsymbol{\theta}_k, \tau_k, z_{ik} | \mathbf{x}) = \sum_{i=1}^{n} \sum_{k=1}^{G} z_{ik} \times \ln\left(\tau_k f_k\left(\mathbf{x}_i | \boldsymbol{\theta}_k\right)\right) \tag{20}$$

Let $\hat{z}_{ik}$ denote the condition expectation of z_{ik} given $\mathbf{x}_i$ and associated parameter values, i.e., $\hat{z}_{ik} = E\left[z_{ik} | \mathbf{x}_i, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_G \right]$, and z_{ik}^* the value of $\hat{z}_{ik}$ at a maximum of Equation (16), which is the conditional probability that the i^{th} PTC belongs to group *k*. Figure 1 illustrates the EM algorithm for clustering via Gaussian mixture models [FRA98]. The EM algorithm alternates between two steps: an "E-step" and an "M-step." During the E-step, values of $\hat{z}_{ik}$ are computed from the data with the current parameter estimates. At the M-step, the complete likelihood for Equation (20) with each z_{ik} replaced by its current conditional expectation $\hat{z}_{ik}$ is maximized with respect to the parameters.

Initialize $\hat{z}_{ik}$

Repeat

M-step: Computer maximum-likelihood parameter estimates given $\hat{z}_{ik}$

$$\hat{n}_k = \sum_{i=1}^{n} \hat{z}_{ik}$$

$$\hat{\tau}_k = \frac{n_k}{n}$$

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{n} \hat{z}_{ik} \mathbf{x}_i}{n_k}$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{\sum_{i=1}^{n} \hat{z}_{ik} \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k\right)\left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k\right)^{\mathrm{T}}}{\hat{n}_k}$$

E-step: Computer $\hat{z}_{ik}$ given the parameter estimates from the M-step

$$\hat{z}_{ik} = \frac{\hat{\tau}_k f_k\left(\mathbf{x}_i \middle| \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k\right)}{\sum_{j=1}^{G} \hat{\tau}_j f_j\left(\mathbf{x}_i \middle| \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j\right)}$$

Until convergence criteria are satisfied.

Figure 1. EM Algorithm for Clustering via Gaussian Mixture Models

The EM algorithm has the following limitations [FRA98]:

- Unless starting with reasonable initial values, the rate of convergence may be slow.
- The number of conditional probabilities associated with each PTC equals the number of components in the mixture. As a result, it is not practical for models with a large number of seasonal factor groups.
- EM breaks down when the covariance matrix corresponding to one or more seasonal factor groups is ill-conditioned, i.e., singular or near singular.

There are two issues in the clustering analysis: the selection of the clustering method, such as those presented in Table 2, and the determination of the number of clusters. Bayesian Information Criterion (BIC), as illustrated in Equation (21), is applicable to find the maximum mixture likelihood [FRA98]:

$$BIC = 2L - r\log(n) \tag{21}$$

where

L = log-likelihood of the model;
r = total number of parameters to be estimated in the model; and
n = number of PTCs.

The number of clusters is not considered an independent parameter for the purpose of computing the BIC. Likelihood cannot be used directly to evaluate a model since the fit of a mixture model to a given data improves as more terms are added to the model. In the expression of BIC, a term is added to the log likelihood to penalize the complexity of the model. Consequently, the BIC allows smaller numbers of groups than the log likelihood does.

2.3 Geographic/Functional Assignment

The method documented in the Bureau of Public Roads' *Guide for Traffic Volume Counting Manual* involves a manual ranking system. Using this method, monthly traffic factors of permanent count stations and the ratio of the AADT to the average weekday traffic of the month are sorted in ascending order [BPR65]. For each month, a group of counters is determined so that the difference between the smallest and the largest factors does not exceed 0.2. The final grouping of counters is manually examined to ensure as many counters as possible fall into the same group in each month.

Bellamy described a subjective classification system for determining the grouping for a site. Four classes were identified as urban/commuter, low flow (< 1000 veh/day) non-recreational rural, rural long-distance, and recreational [BEL78]. Sharma proposed a method to classify rural roads based on trip purpose and trip length information collected from past origin-destination surveys by Alberta Transportation [SHA83, SHA86]. Traffic counters were first grouped according to their monthly traffic patterns by hierarchical grouping. For counters in the same group, the daily and hourly traffic variations for the months of May to August were then examined. Based on the daily traffic patterns collected in 1978 or 1977 from a total of 45 counter sites, the following five predominant road uses were identified [SHA83]: commuter (COM), commuter-recreational (CR), commuter-recreational-tourist (CRT), tourist (TOUR), and

highly recreational (HREC). The following three typical patterns of hourly volumes were also identified: commuter pattern, partially commuter pattern, and non-commuter pattern. Trip purpose data and trip length data from past origin-destination survey were then utilized to investigate the effects of travel behaviors on counter grouping. Trip purpose data were used to verify the temporal volume variations, which were categorized into the following two groups: work-business and social-recreation. Cumulative trip length distribution information was used to classify roads for mainly serving regional, interregional, or long-distance travel. Seven road classes were defined: commuter, commuter-recreational, commuter-recreational-tourist, rural long distance, non-recreational low volume, high recreational, and special. The same procedure was used to examine the data from 52 sites in Alberta and the grouping was tested on the data from 28 sites in Saskatchewan, Canada [SHA86]. Eight road classes were consequently defined: regional commuter, regional recreational and commuter, interregional, long distance, long distance and recreational, highly recreational, rural commuter and business, and special. Faghri and Hua classified roads into urban/rural, recreation/non-recreational, and recreational–arterial/otherwise based on their physical and functional characteristics [FAG95]. To estimate the number of automatic traffic recorders (ATRs) needed, Faghri et al. classified count sites into four categories based on the value of monthly coefficient of variation (MCV): urban (MCV < 10%), rural (10% ≤ MCV < 25%), recreational (25% ≤ MCV < 35%), and predominantly recreational (MCV > 35%) [FAG86]. Such classification of traffic characteristics, however, is difficult to obtain for large urban areas due to the dispersion and mixing of different types of activity centers, making it unlikely that a particular type of trips will be the dominant traffic on a given road.

Ritchie proposed a statistical framework to analyze statewide traffic count data [RIT86]. This approach incorporated seasonal effect on traffic volumes by first stratifying highway system according to geographic region and functional classification. The strata with similar seasonal patterns were combined. Using the data collected from 1980 to 1984 in Washington, seven groups were obtained: rural interstates, urban roads, other rural roads in the northeastern, southeastern, northwestern, and southwestern parts of Washington, and central mountain passes. The following regression model was then calibrated to estimate seasonal factors for each group:

$$\text{seasonal Factor}_i = \frac{\text{AADT}}{\text{VOL}_i} = \beta + \varepsilon \tag{22}$$

where

VOL_i = average 24-hour short-count volumes calculated from the 72-hour Tuesday-Thursday counts for month *i;*

ε = error term whose variance was considered as a constant; and

β = regression coefficient, which was interpreted as the estimated seasonal factor for a specific factor group for a given month.

Delaware Department of Transportation (DelDOT) utilized the procedure suggested in the FHWA Traffic Monitor Guide (TMG) and categorized permanent count stations according to their monthly coefficient of variations (MCVs) into the following four groups: urban group for MCVs less than 10%, rural group for MCVs between 10% and 25%, recreational group for MCVs between 25.1% and 35.1%, and predominantly recreational group for MCV greater than

35.1% [FAG96]. The MCVs were determined using the following formula where M_i is the monthly AADT:

$$\text{MCV} = \frac{\sqrt{\frac{1}{2}\sum_{i=1}^{12}(M_i - \text{AADT})^2}}{\text{AADT}} \tag{23}$$

Virginia Department of Transportation (VDOT) applied an approach to factor short-term vehicle classification counts by simultaneously considering seasonal and weekly traffic variations [WEI96]. VDOT first classified vehicles into the following five groups:

1. Four-tire vehicles (Classes 2 and 3)
2. Buses (Class 4)
3. Other six-tire, two-axle vehicles (Class 5)
4. Other single-unit vehicles with three of more axles (Class 6 and 7)
5. Combination trucks (Classes 8-13)

Two sets of 84 combined-month-and-day-of-week (CMDW) factors, which were developed for each day of the week and month of the year, i.e., 7 days × 12 months, were calculated for vehicle groups 1 and 5. For vehicle groups 2 through 4, seven day-of-week factors and 12 separate monthly factors were developed and applied in pairs to reduce the mean absolute error (MAE). The error was the average of the absolute values of the percent differences between the estimated and actual AADT.

Kentucky Transportation Cabinet applied a similar approach to that of VDOT to factor short-term vehicle classification counts [STA97]. Eighty-four CMDW factors were developed for four different types of roadways, i.e., rural Interstates and parkways, urban Interstates and parkways, rural non-Interstates and non-parkways, and urban non-Interstates and non-parkways, for each of the 15 vehicle types. The preliminary validation showed that more accurate AADT estimates were obtained when each vehicle type was factored alone and then estimates for different vehicle types were added to obtain the overall AADT for a given roadway segment.

2.4 Regression Analysis

Regression techniques may be used as a tool to analyze the relationship between seasonal variations in traffic volume and some predictors [FAG95]. Variables used generally include those that represent the physical and functional characteristics or their combinations. Dummy regressors are used to represent these characteristics as the "yes/no" type of variables. The regression model is commonly defined in a linear form:

$$f_{sm} = \alpha_{0m} + \alpha_{1m}x_1 + \alpha_{2m}x_2 + \ldots \tag{24}$$

where

f_{sm} = seasonal factor in month m; and
x_i = dummy variable that takes the value of 0 and 1 (i = 1, 2 , …).

Faghri and Hua concluded that urban/rural, recreation/non-recreational, and recreation-arterial/otherwise variables were statistically significant and could provide better results than cluster analysis [FAG95].

Regression analysis was also often used to estimate AADT directly to avoid the use of seasonal factors. For example, the following simple linear regression equation was used to estimate AADT [LAM00]:

$$y = A + Bx \tag{25}$$

where

y = estimated AADT;
x = short period count at the selected station; and
A, B = regression coefficients.

Erhunmwunsee compared AADTs estimated from multiple regression analysis with those from the Philips and Blake's method based on the traffic data from the City of Milwaukee's 24 continuous count stations [ERH91]. A total of 12 stations were randomly selected as the fitting samples and the remaining stations were validation samples. The regression analysis equation was defined as follows:

$$y = B_0 + B_1 \times F_{ij1} + B_2 \times F_{ij2} + \ldots + B_n \times F_{ijn} \tag{26}$$

where

y = estimated AADT;
B_i = regression coefficients (i = 1, 2, … n); and
F_{ij} = short period count at station i on day j on month 1, 2, … n.

It was determined that the period with its midpoint centered at 3 PM was the best period in a day to begin a short-term count and that the best month to conduct short-term counts was April, followed by June and October. Erhunmwunsee also concluded that the regression method produced the better AADT estimates than the Philips and Blake's method.

Seaver *et al*. proposed a statistical procedure utilizing principal component analysis, multivariate regression, regression clustering, and multiple regression analysis to model ADT on rural local roads [SEA00]. Data collected from 80 randomly selected counties in Georgia were utilized for model development. The procedure had the following steps:

1. Apply principal component analysis to identify p principal components (y_1, y_2, …, y_p) from n initial independent variables (x_1, x_2, …, x_n) for each paved (Road Type 4) and unpaved (Road Type 5) rural roads in the metropolitan statistical area (MSA) and non-MSA.
2. Apply multivariate regression to find the principal variables from the n initial independent variables (x_1, x_2, …, x_n) that are correlated with the principal components (y_1, y_2, …, y_p) identified in the first step.

3. Apply regression clustering to determine strata for each road type in both MSA and non-MSA by using the ADT in a county as the dependent variable and the principal variables identified in Step 2 as the regressors.
4. Perform a multiple regression on the data within each cluster.

The advantage of this method is that all independent variables used in the procedure for developing the models were obtained from the U.S. census. The time and cost for obtaining the data were subsequently reduced. The disadvantage, on the other hand, is that census may not be up-to-date and data verification is needed. The statistical procedure proposed by Seaver *et al*. may be applicable to grouping PTCs into seasonal clusters for AADT estimates, provided that data for the independent variables at PTC level are available.

Zhao and Chung performed various multiple liner regression analyses to investigate factors affecting AADT estimates in Broward County, Florida [ZHA01]. Geographic information system (GIS) technology was utilized to compile intensive land-use and accessibility measures. Four models were calibrated after outliers were removed. Two variables, i.e., functional classification and number of lanes, were found to be the most significant predictors for estimating AADTs. Other land-use variables, including direct access to expressway, employment size in the buffer area around a given count station, distance to spatial mean centers of population, and regional accessibility to employment centers, were also found to be significant.

Davis applied the weighted least-squares regression to calibrate the following model for traffic counts [DAV97]:

$$y_t = \mu + \sum_{i=1}^{12} \Delta_{t,i} m_{k,i} + \sum_{j=1}^{7} \delta_{t,j} w_{k,j} + \varepsilon_t \qquad (27)$$

where

y_t = natural logarithm of the traffic count on day t;
μ = expected log traffic count on a typical day;
$\Delta_{t,i}$ = 1, if the count was made during month i (i = 1, …, 12), and 0 otherwise;
$m_{k,i}$ = correction term for month i, characteristic of factor group k (k = 1, 2, or 3);
$\delta_{t,j}$ = 1, if the count was made on day-of-week j (j = 1, …, 7), and 0 otherwise;
$w_{k,j}$ = correction term for day-of-week j, characteristic of factor group k; and
ε_t = random error (residual).

After eliminating missing and imputed data from the traffic data collected in 1992, a total of 50 ATRs that were classified into three factor groups by the Minnesota Department of Transportation (MnDOT) personnel were included in the model development. The mean-value (μ), monthly ($m_{k,i}$), and day-of-week ($w_{k,j}$) terms were estimated using re-weighted least squares in MINITAB with the following procedure iteratively:

- Estimate mean, monthly, and day-of-weeks parameters via the GLM procedure in MINITAB;

- Compute the residual variance for each ATR in a group given the current regression parameter estimates; and
- Use the variance estimates to compute separate weighting vectors for each ATR.

The monthly and day-of-week terms were constants for all ATRs within a specific factor group k, but each ATR in the factor group was allowed to have its own mean-value parameter μ. The weighted least squares approach in the MINITAB's GLM procedure was chosen due to the heteroscedasticity caused by ATRs' different day-to-day variances. The residuals, i.e., ε_t's, were further validated with their temporal dependency, and the following seasonal, multiplicative autoregressive model was obtained:

$$\varepsilon_t = \Phi_1 \varepsilon_{t-1} + \Phi_7 \varepsilon_{t-7} - \Phi_1 \Phi_7 \varepsilon_{t-8} + a_t \tag{28}$$

where

a_t = independently and identically distributed normal random variables with mean equal to 0 and common variance; and

Φ_1, Φ_7 = autoregressive coefficients.

Once the parameters in the above autoregressive model were estimated, i.e., $\hat{\Phi}_1$ and $\hat{\Phi}_7$, the residuals, i.e., $\varepsilon_t - \left(\hat{\Phi}_1 \varepsilon_{t-1} + \hat{\Phi}_7 \varepsilon_{t-7} - \hat{\Phi}_1 \hat{\Phi}_7 \varepsilon_{t-8}\right)$, were validated to confirm that the residuals were not significantly autocorrelated and would pass the goodness-of-fit test of being normally distributed.

2.5 Artificial Neural Networks

Artificial neural networks (ANNs) are computing techniques that attempt to simulate the workings of the human brain. It is known that ANNs are superior to traditional computing techniques in solving pattern classification problems due to their unique properties [FAG95]:

- Ability to deal with incomplete input information;
- Ability to deal with noisy input data; and
- Ability to learn and associate patterns from historical data.

The ANN models consist of many simple processing elements, i.e., neurons, with dense parallel interconnections. They may be classified according to various criteria, such as their learning methods (supervised versus unsupervised), architectures (feed-forward versus recurrent), output types (binary versus continuous), node types (uniform versus hybrid), implementations (software versus hardware), connection weights (adjustable versus hardwired), operations (biologically motivated versus psychologically motivated), etc. [JAN97]. In ANNs, feed-forward means the output of each processing element generally propagates from the input side to the output side. If there is a feedback link that forms a circular path in a network, the network is called recurrent.

Training and testing are the two stages in the development of an ANN model. During the training stage, an inductive learning principle is used to learn from a set of examples called a training set. Several neural network learning schemes, including supervised learning and unsupervised learning, have been developed. A supervised learning ANN is first trained by a

selected algorithm to learn from the AADTs collected at permanent count stations. The trained ANN may then be used to estimate AADTs at short count stations. Consequently, unlike traditional method of estimating AADT from sample volume counts, determining ATR factor groups according to similarities in their temporal traffic variations and then assigning each short count station to one of the established factor groups is no longer a priori. Unsupervised learning ANNs may be trained without any information of the desired output to determine factor groups after frequently occurring traffic patterns are recognized. The following sections provide a brief introduction for supervised and unsupervised learning methods and their applications to grouping traffic patterns and/or AADT estimation.

2.5.1 Supervised Learning

Supervised learning involves providing an ANN with "examples" that consist of inputs and the corresponding outputs. The learning algorithm attempts to adjust the weights of the connections between neurons to produce the desired output. As a result, such networks are also referred to as mapping networks. During the mapping process, the error in the output is propagated back to the previous neurons by adjusting the weights of the connections. This is called the back-propagation (BP) method for propagating the error, or known as the generalized delta rule (GDR). Figure 2 illustrates the architecture for supervised learning and back-propagation neural networks where the target is the desired output. The process begins by assigning weights with small random values and terminates when either the maximum number of iterations is reached or the sum of absolute error (SAE) is reduced to an acceptable value.

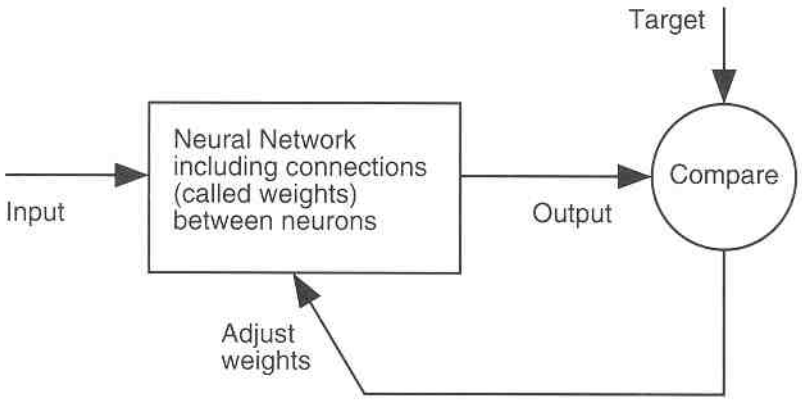


Figure 2. Neural Network Architecture

The multi-layered feed-forward network is probably the most commonly used model for estimating AADT. Sharma *et al*. investigated the traffic volume data from 63 ATR sites located on the regional and rural roads in Minnesota using a multi-layered, feed-forward, back-propagation, and supervised learning approach [SHA99]. The data were collected during a period between May and August in 1993. The model consisted of three layers of neurons, i.e., one input layer, one output layer, and one hidden layer for feeding data from the input layer to the output layer. The input was the hourly volumes of vehicles included in a sample counting program divided by the sample average daily traffic (SADT), which was simply the total volume for one or more short-period traffic counts in the sample divided by the number of sample days.

Therefore, the number of neurons in the input layer was equal to the total number of hourly volumes. The hidden layer had half of the number of neurons in the input layer, and the output layer only contained one neuron, which gave the estimated value for an AADT factor. The actual AADT factor for the output layer was defined as follows:

$$\text{Actual AADT factor} = \frac{0.25 \times AADT}{SADT} \tag{29}$$

The estimated AADT was calculated using the following equation:

$$\text{Estimated AADT} = 4 \times (\text{SADT} \times \text{output factor from ANNs}) \tag{30}$$

The learning cycles were set at 25,000. The results from the neural networks were compared with those from the traditional hierarchical grouping method proposed by Sharma and Werner [SHA81]. Although the comparison from their study indicated that the errors from the neural network model were larger than those from a traditional grouping method, the authors argued that short-period count sites could not be assigned 100 percent correctly to one of the factor groups in practice. As a result, the neural network approach would be a better alternative to estimate AADT since it would not require classifying permanent count stations to groups and then assigning sample count sites to their associated PTC groups.

Sharma *et al*. reached similar conclusions as those in [SHA99] regarding the accuracy of the AADT estimates on low-volume roads using the traditional factor approach and ANNs [SHA00, SHA01]. The supervised learning ANNs structure, same as that implemented in [SHA99] with the multi-layered, feed-forward, and back-propagation design was again utilized. The traffic volume data collected from 55 ATR sites located on the rural roads in Alberta, Canada in 1996 were investigated. The low-volume roads referred to those for which the AADT volumes were between 120 and 999 vehicles. Sharma *et al*. concluded again that the factor approach produced better AADT estimates than ANNs if the ATR sites were grouped appropriately and the sample sites were correctly assigned to their associated groups.

Lam and Xu implemented a multi-layered feed-forward, back-propagation neural network that consisted of one input layer, one output layer, and one hidden layer to group the traffic flow data collected in 1991 from 13 count sites in Hong Kong [LAM00]. Different lengths of counts, i.e., four hours, six hours, eight hours, 10 hours, 12 hours, 14 hours, and 16 hours, each associated with several starting times in a day, were investigated. The sum of absolute percentage errors (SAE) from the PTCs included in the study was calculated using the following equation for both methods for the 13 count stations:

$$SAE = \sum_{i}^{13} \left| Error_i(\%) \right| \tag{31}$$

where Errori(%) is the percentage error between the estimated and actual AADT at the *i*th PTC. An effectiveness index (*Eff*) was defined to measure the effect of the extra counting time under the assumption that the cost of traffic counts was proportional to the count duration:

$$Eff = \frac{RSAE}{ETLC} \tag{32}$$

where

RSAE = reduction in SAE; and
ETLC = amount of extra time length of count.

By comparing SAEs, Lam and Xu concluded that the neural network approach consistently performed better than the regression analysis approach in estimating AADT. The 12-hour count period was found to be the most accurate period for AADT estimation because of the minimum SAE. However, the 8-hour count was the most effective period of count with the highest *Eff* value of 5.41.

Lingras *et al*. applied a time-delay neural network (TDNN) and an autoregression (AR) model to forecast daily traffic volumes at 78 PTC sites in Alberta, Canada [LIN00]. To simplify the analysis, the PTC sites were first classified into the following five types of road groups:

1. Highly recreational
2. Regional recreational
3. Long distance
4. Urban commuter
5. Regional commuter

The method suggested by Sharma and Werner [SHA81] was used to determine different groups of road classifications based on the traffic data collected in 1993. After road types were determined, one PTC site each from groups 1, 2, and 3 and three PTC sites each from groups 4 and 5 were selected. Only PTC sites with continuous traffic data from 1989 to 1993, inclusive, were selected. Their traffic data collected from 1989 through 1992 were then used to train the TDNN and calibrate AR model for each classification group. These models were subsequently tested using the data collected from these selected PTC sites in 1993. Daily traffic volumes of the previous 13 days (i.e., x_1, x_2, …, x_{13},) were defined as the independent variables or input variables to predict traffic volume for the following day (x_{14}). The AR equation is shown below.

$$x_{14} = \sum_{i=1}^{13} a_i x_{14-i} + e_{14} \tag{33}$$

The TDNN had 13 input nodes corresponding to the previous 13 daily traffic volumes and one output to predict the traffic volume. The average and maximum percentage errors (between the predicted and the actual traffic volumes) as well as the 50^{th}, 85^{th}, and 95^{th} percentile errors from the cumulative frequency distributions were used as the model performance measures. Lingras *et al*. concluded that TDNN models produced better predictions than AR models for all the five road groups since all of the error measures were smaller with the neural network approach.

Theoretically, if the number of neurons in the hidden layer is large enough, supervised learning ANNs will be able to approximate closely any complicated non-linear function. Current practices, however, utilize the ANN paradigm designed with one hidden layer to reduce the

intensive computing efforts in the training process. Consequently, the performance of supervised learning neural networks on estimating AADT has not been truly explored.

2.5.2 Unsupervised Learning

Unsupervised learning, or learning without supervision, is an approach that extracts features or regularities from presented patterns without any information of the desired output [JAN97]. ANNs with unsupervised learning update weights only on the basis of the input patterns and are trained to respond to frequently occurring patterns. The following sections describe the unsupervised learning paradigms for competitive learning and the Kohonen self-organizing feature map.

2.5.2.1 *Competitive Learning and ART1*

In competitive learning ANNs, the number of output units is equal to the number of clusters into which the data are divided. The weights of the neural connections are updated according to the competitive, or winner-take-all, learning rule. Competitive learning ANNs have two disadvantages. One is that the number of classification clusters has to be specified before the learning proceeds and the model lacks the capability to add new clusters when necessary. In other words, competitive learning classifies a given pattern into exactly one of the mutually exclusive classes that are predetermined. The other is that response to the same input pattern may differ on each successive presentation of that input pattern and the winning unit that responses to a particular pattern may continue to change during training. This is usually referred to as the stability-plasticity dilemma. Such unstable learning in response to prescribed input is due to the learning that occurs with other intervening inputs. Consequently, the network adaptability, or plasticity, enables prior learning to be erased by more recent learning in response to a wide variety of input environments. As a solution to the dilemma, Carpenter and Grossberg proposed the ART1 architecture that was capable of recognizing patterns from arbitrary binary input patterns [CAR88]. The ART1 neural network is a paradigm of adaptive resonance theory (ART) that processes binary patterns in which each element of input vector takes only a value of 0 or 1. The ART1 learning scheme is also capable of creating new clusters when needed.

Faghri and Hua applied the ART1 neural network to group 29 ATR stations in Delaware with traffic data collected from 1985 through 1989 [FAG95]. ART1 had only one layer of processing units. ART1 ANNs set up certain categories for the input and classified the input pattern into a proper category. If an input pattern did not match any existing categories, the network would create a new category for it. The ratio of a MADT to the corresponding AADT, i.e., V_0, for a given PTC was first converted using the following formula:

$$V_n = \frac{V_0 - V_0^{\min}}{V_0^{\max} - V_0^{\min}} \tag{34}$$

where

V_n = conversion result for the ratio of MADT to AADT;
V_0 = ratio of MADT to AADT;
V_0^{max} = maximum value of the MADT to AADT ratio; and

V_0^{min} = minimum value of the MADT to AADT ratio.

The 12 new ratios corresponding to the 12 months in a year were then converted to binary numbers and entered into each column of a 10 × 12 matrix. This matrix was used as an input to the ART1 ANNs for the traffic pattern obtained from a given PTC. Some accuracy was lost due to rounding because each MSF was represented by a 10 × 1 vector. This loss of accuracy was considered insignificant and ignored in the study. A value of 0.83 was determined as the vigilance factor after a few pre-designated count sites were correctly classified into proper categories. The results from the ART1 method were compared with those obtained from both cluster and regression analyses. While four seasonal categories were produced by all three methods, they differed in the way that the ATRs were grouped. Cluster and regression analyses created categories of urban, rural, recreational arterial, and recreational collector, while the ANN created categories of urban or interstate, rural arterial, rural collector, and recreation. There were only two ATR stations whose categories were not determined by the ART1 method. For at least five ATR stations, the groups changed from year to year and from method to method.

The following equation was used to measure the comprehensive performance of the three methods in estimating seasonal factors:

$$average_{type} = \frac{1}{12}\sum_{\forall j} err_{type}(j) = \frac{1}{12}\sum_{\forall j}\sum_{\forall i}\left[sf_{type}(i,j) - sf_{act}(i,j)\right]^2 \quad (35)$$

where

$average_{type}$ = average error for method *type*;
$err_{type}(j)$ = dissimilarity between estimated and actual seasonal factors in month *j* for method *type*;
$sf_{type}(i, j)$ = estimated seasonal factor for method *type* at *i* ART in *j* month; and
$sf_{act}(i, j)$ = actual seasonal factor at *i* ART in *j* month.

By comparing the average errors from the three grouping methods, Faghri and Hua concluded that the neural network method outperformed the cluster and regression methods. The results indicated that ART1 networks had the ability to organize inputs into their natural groups as well as the capability of weeding out random seasonal fluctuations in the input patterns.

2.5.2.2 *Kohonen Self-Organizing Feature Map*

Kohonen self-organizing networks, also known as Kohonen feature maps or topology-preserving maps, are another competition-based network paradigm for data grouping. The learning procedure of Kohonen feature maps is similar to that of competitive learning ANNs. However, in addition to updating the weights for the winning units, all the weights in a neighborhood surrounding the winning units are updated as well. The network consists of two layers, i.e., input and Kohonen layers. The network receives the input vector as a given pattern. If the pattern belongs to the k^{th} group, the k^{th} unit in the Kohonen layers will have an output value of one while the other neurons will have a value of zero.

Lingras compared the classification groups from Kohonen unsupervised learning ANNs and with those from a hierarchical grouping method using data collected from 72 PTC sites in Alberta, Canada [LIN95]. Five seasonal categories were specified for Kohonen ANNs. The number of iterations was set to 100 since grouping stabilized after presenting the training set to the ANN 100 times. The findings included that the Kohonen ANNs produced results that were similar to the hierarchical groping method. As a result, ANNs could be used to substitute the statistical techniques for grouping of traffic patterns. Moreover, Kohonen ANNs updated the weights on the connections only when complete patterns were presented. For incomplete patterns, the ANNs could find the categories using the least mean-square error or other similarity measures. This feature enabled Kohonen ANNs to classify incomplete monthly traffic patterns.

2.6 Genetic Algorithms

Genetic algorithms (GAs), originally called genetic plans, have received a great deal of attention because of their potentials to solve optimization problems [SAK02]. The GA technique is a stochastic search process based on the mechanism of natural selection and genetics. In a GA, problem solutions are represented as chromosomes, which are made up by genes. Starting with an initial population of individuals, i.e., chromosomes, genetic operators are applied to evolve the population by producing successively new populations with improved "fitness" of the individuals. Each iteration produces a new generation of solutions. For any given generation, each individual in the population is evaluated using some measure of fitness, usually the objective function in an optimization problem. Genetic operators, such as selection, reproduction, crossover, and mutation, are then used to create the next generation of the population. Selection is to select individuals from the current population based on their fitness values. Reproduction involves applying crossover and mutation operators to some of the selected individuals to produce a new generation whose overall fitness should improve over the previous generation. The crossover operator selects individuals from the population at random and exchange portions of the genes to produce new individuals, while the mutation operator randomly alters one or more genes of a selected individual. The process continues until the termination condition is satisfied, which is either the best fitness value of the population stops to improve or a prescribed number of iteration is exceeded. The general framework of genetic algorithms is presented in Figure 3, where $P(t)$ denotes the population at generation t.

```
begin
        t := 0
        initialize P(t)
        evaluate P(t)
        while (not termination condition) do
        begin
                t := t + 1
                select P(t) from P(t - 1)
                alter P(t)
                evaluate P(t)
        end
end.
```

Figure 3. General Framework of Genetic Algorithms

GA-based methods have several advantages:

- GA formulations do not require the calculation of gradient matrices or other higher order derivative matrices or their approximations.

- A GA-based solution method directly operates its search process, e.g., transformation through genetic operators and selection based on fitness. Therefore, there is no need to formulate a system of governing equations that represents or simulate mathematically the relationship between various parameters and unknowns. This is particularly attractive for practical applications where it is difficult to establish mathematical formulation to accurately and effectively simulate complex problems.

- Constraint conditions are relatively easy to incorporate into a GA solution process. Constraint conditions may be simply defined as a part of the environmental conditions or by assigning large penalty numbers to individuals that violate certain constraints thus reducing their surviving possibility in the selection process. This may be especially suitable to problems where constraints are complicated and cannot be properly defined.

GAs have been a very active research field for the past several decades and results have been widely used in various application fields. However, GAs also have two main disadvantages:

- GAs are stochastic algorithms whose search methods are based on the natural evolution principle. Although a sufficiently large number of "individuals" may result in a nearly optimal solution to an optimization problem, the GA technique does not guarantee global optimal solutions.

- GAs may require extremely large amount of computer CPU time when dealing with large-scale problems.

Lingras utilized a GA to group PTCs and compared the classifications with those from the traditional hierarchical grouping method developed by Sharma and Werner [LIN01]. The monthly traffic patterns collected between 1987 and 1991 from PTC sites on Alberta highways were used. The number of genes in a chromosome was set to equal to the number of seasonal patterns that needed to be classified. Each chromosome corresponded to a classification scheme. A gene was randomly assigned with an initial value between 1 and *m*, where *m* is the desired number of groups. Solutions of two to 15 factor groups with the following object function were investigated:

$$\Delta_1 \Big/ \Delta_m = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d\left(P_i, P_j\right)}{\sum_{i=1}^{m}\sum_{x_j, x_k \in X_i} d\left(x_j, x_k\right)} \tag{36}$$

where

Δ_1 = maximum possible within-group error;

Δ_m = sum of within-group error for m groups of seasonal patterns;
P_i = seasonal traffic pattern i;
$d()$ = a distance function to measure the dissimilarity between patterns;
x_j = seasonal traffic pattern j in factor group X_i; and
n = total number of seasonal patterns.

The behaviors of both GAs and hierarchical methods were also compared for 20, 30, 40, and 50 groups. Galib, a program available at http://lancet.mit.edu/ga/, was used to perform the GA analysis. The classification schemes for different numbers of groups with the highest values of Δ_1/Δ_m from 1,000 generations of evolution were compared with those from the traditional hierarchical clustering approach. The results indicated that the hierarchical grouping method performed better when the number of groups was greater than 14. However, GAs performed better when the number of groups was less than nine. Since the initial grouping patterns were randomly assigned, the results were verified by repeating the experiment for five factor groups 22 times. The within-groups errors varied between 680 and 730, which were consistently and significantly lower than the hierarchical grouping error of 861. The genetic approach was also applied with different numbers of generations ranging from 100 to 1,000 with an interval of 100 for five factor groups. The results showed that the GAs errors were less than the hierarchical groping error after 400 generations.

2.7 Assignment of Count Sites

There is considerable vagueness in the current practice of assigning count sites to seasonal factor groups. Currently, assigning short-count sites to factor groups and determining the precision of short count estimates are generally accomplished by considering the physical proximity of short count sites to a PTC site and based on engineering judgment [TMG01]. If the true factor group for a site is known, it was reported that traditional short-counts could provide estimates of mean daily traffic with the PI95 (precision achievable with 95% confidence) between 10 and 23 percent [DAV96a]. Inappropriately assigning a site to a factor group may result in a great decline in precision.

North Carolina Department of Transportation (NCDOT) implemented a data management system developed with GIS to assign each short count site to one of the seven seasonal groups based on the most recent data at that site [MCD99]. Short count stations that had at least three or four 48-hour sampling traffic counts available were used to identify the seasonal group that was highly correlated over these short count stations' day and month variations. In other words, statistical correlation and their associated p-values were used to determine the best seasonal group for a given short count site.

Davis and Guan employed Bayesian theorem to assign a given site to a known seasonal factor group with the highest posterior probability [DAV97]. The probability was defined as follows:

$$\text{Prob}[\text{site} \in \text{G}_\text{k} \mid z_1, \ldots, z_N] = \frac{f(z_1, \ldots, z_N \mid G_k)\,\alpha_k}{\sum_{l=1}^{n} f(z_1, \ldots, z_N \mid G_l)\,\alpha_l} \tag{37}$$

where

$f(z_1, \ldots, z_N)$ = a likelihood function measuring the probability of obtaining the count sample had the site belonged to a given seasonal factor group;
$z_1, \ldots, z_N$ = a sequence of N daily traffic counts at a short-count site;
$G_1, \ldots, G_n$ = a total of n different factor groups; and
α_k = probability that the given site belongs to G_k prior classification.

The prior classification probability, i.e., α_k, was assumed to equal to $1/n$, indicating complete prior uncertainty as to which group a short count site belonged. The linear regression model described in Section 2.4 was used as the likelihood function in the posterior classification probabilities, which is again given below:

$$y_t = \mu + \sum_{i=1}^{12} \Delta_{t,i} m_{k,i} + \sum_{j=1}^{7} \delta_{t,j} w_{k,j} + \varepsilon_t \tag{38}$$

It was further assumed that $\varepsilon_1, \ldots, \varepsilon_N$, were normally distributed random errors with a mean value of 0 and a covariance matrix $\sigma^2 \mathrm{V}$, where σ^2 was the common unconditional variance of y_t, and V is a $N \times N$ matrix of correlation coefficients such that the element in row s and column t, $V_{s,t}$, was the correlation coefficient for y_s and y_t. The approach was developed based on the assumption that short-term count sites should be assigned to one of the seasonal factor groups that had a similar monthly and daily variation pattern. The model was validated using data from 48 ATR stations for 1991 and 50 for 1992. This data-driven approach was shown to be able to produce mean daily traffic estimates that were near ±20 percent of actual values based on 14 well selected sampling days from particular months and days of the week. Although the method did not provide significant improvements in precision over what may be achieved when the appropriate seasonal factors were known, reliance on subjective judgment was reduced in the process. A potential problem with the Bayesian assignment approach proposed by Davis and Guan, however, is that longer period of data collection at short count sites is needed. The approach is also complicated and time consuming to implement.

2.8 Other Issues

Two important issues need to be considered in seasonal factor modeling, i.e., data quality and precision. Missing or erroneous data due to machine failures, system errors, or vandalism are often encountered in the analysis of traffic data. Data imputation generally refers to editing and correction of data that are missing or inconsistent. The following sections first discuss the effects of data imputation and available approaches to identify and correct such data. The precision analysis recommended in TMG to validate the number of count stations within a given factor group is also described.

2.8.1 Data Imputation

There have been studies to investigate the effect of missing traffic data on AADT estimates. However, to our best knowledge, previous research efforts have not yet considered the effect of missing data on seasonal factor estimates and consequently no conclusions have been reached.

Using continuous traffic data collected in 1994 from 21 permanent count stations, eight classification stations, and six weigh-in-motion stations in Florida, Wright *et al*. investigated the effect of the following three levels of missing data on AADT estimates [WRI97]:

- 5% of days of data missing at random;
- 20% of days of data missing at random; and
- 50% of days of data missing at random.

A given percentage (i.e., 5%, 20%, or 50%) of days in 1994 were randomly selected and excluded from the calculations of AADT and CV for a specific count site. This random sampling procedure was repeated for 1,000 times, each producing an AADT estimate and its associated CV. The simulated AADT and CV, i.e., the average value of the 1,000 AADTs and CVs, were then compared with the respective statistics from the full data set for each count site. The same process was applied to randomly eliminate weekly data up to 8 consecutive weeks. The authors concluded that random missing data did not significantly bias the estimation of AADT. The authors also examined the effect of holidays and special days on AADT under the following three conditions:

- All days of data used;
- Data with all specific holidays removed; and
- Data with all holiday period days removed.

Wright *et al*. concluded that the effect of holidays and special days was negligible on overall AADT estimates. No further investigation was performed to examine the effects of missing or erroneous data on seasonal factors.

There are at least two approaches that may be potentially suitable for data imputation. They are factoring approach and time series analysis. For example, the Pennsylvania Department of Transportation (PennDOT) implemented a factoring approach to perform data imputation. The routine used in the PennDOT's automatic traffic data management system identifies missing or erroneous data when any of the following four criteria are met [CHU98]:

1. Volumes at 1 a.m. are larger than volumes at 1 p.m.;
2. Same volumes are collected for four consecutive hours or longer;
3. Zero volumes are collected for eight consecutive hours or longer; and
4. Other data are missing.

Using data from a particular count site, a lookup table with values determined according to the proportions of volumes at a given time of day and day of the week for each vehicle class will be created. The values from the lookup table are suggested to replace the missing or erroneous data. Time series analysis, such as autoregressive integrated moving average (ARIMA) multivariate models and seasonal exponential smoothing models, may also be implemented to repair imputed data. To be more concise, ARIMA models are linear estimators regressed on past values of the modeled time series (the autoregressive terms) or past prediction errors (the moving average terms) [WIL98]. On the other hand, seasonal exponential smoothing models are linear estimators that place exponentially decayed weights on past values. The rate of this decay is

determined by a set of smoothing parameters that are set to minimize the mean squared error of the one-step forecasts.

Kopanezou and Trivellas considered seasonal and weekly periodical variations and changes during holidays or special event days in their time series analysis to forecast daily traffic volumes [KOP89]. Traffic volumes collected in 1985 from a permanent count station was used for the model development purpose. The following ARIMA multivariate time series model was developed:

$$Z_t = f(x_t; w) + N_t \tag{39}$$

$$f(x_t; w) = w_1 x_{1t} + w_2 x_{2t} + w_3 x_{3t} + w_4 x_{4t} \tag{40}$$

$$\left(1 - B^7\right) N_t = \left(1 + \theta_1 B + \theta_2 B^2 + \theta_3 B^3\right)\left(1 + \theta B^7\right)\alpha_t \tag{41}$$

where

Zt	=	estimated daily traffic on a given day t;
$f(x_t; w)$	=	autoregressive function in the ARIMA model;
N_t	=	moving average function in the ARIMA model;
x_{1t}	=	a dummy variable: 1 if day t falls in the months of January, February, March, or October, and zero otherwise;
x_{2t}	=	a dummy variable: 1 if day t falls in the months of April, May, September, November, and December, and zero otherwise;
x_{3t}	=	a dummy variable: 1 if day t falls in the months of June, July, and August, and zero otherwise;
x_{4t}	=	a dummy variable: 1 if day t is a holiday or special even day;
B	=	backward-shift operator; and
α_t	=	white noise, which is identically and normally distributed with mean zero and variance σ^2.

The classification of the monthly factor was obtained by utilizing the least significant difference (LSD) multiple comparisons methodology. The MADT was first calculated and sorted in ascending order. The hypothesis test using the Fisher multiple pair-wise comparisons method was subsequently performed to identify the monthly groups. The nonlinear least square routine from the statistical package RATS was used to estimated the time series model parameters, i.e., w_1, w_2, w_3, w_4, θ_1, θ_2, θ_3, and θ. The model was then evaluated using the data collected from January through April at the same PTC sites in 1986. The results indicated that time series techniques could provide highly accurate and inexpensive short-term forecasts.

2.8.2 Precision Analysis

Precision analysis for seasonal factors is to determine if the number of ATR locations in each factor group is adequate to achieve the desired level of accuracy for the composite group factors [TMG01]. By assuming the permanent count stations in a given seasonal group being randomly selected, the confidence intervals may be estimated as:

$$d = t_{\alpha/2,n-1} \frac{c}{\sqrt{n}} \tag{42}$$

where

d = precision as a percentage to the mean factor for a given month;
c = coefficient of variation of the seasonal factors in a group;
n = number of ATRs in a given group; and
$t_{\alpha/2,n-1}$ = t statistic at 100(1 - α) percent confidence interval and n – 1 degrees of freedom.

The reliability levels recommended by TMG are 10 percent precision and 95 percent confidence interval for each individual seasonal group, excluding recreational groups. The number of ATRs needed is usually five to eight per seasonal group but may vary from group to group.

French *et al.* conducted precision analysis using seasonal factor data collected in 1998 and 1999 from 63 permanent count stations operated by PennDOT and found only one deficient group out of the ten traffic pattern groups used by the agency [FRE01]. Currently, there is no explicit guideline from the TMG for computing growth factors to estimate AADTs and consequently no precision analysis is recommended [TMG01]. PennDOT applied the following procedure, which is similar to the approach for seasonal factor analysis, to estimate the number of sites needed for each of the 42 growth factor categories [FRE01]:

- Calculate growth factor for each ATR in each growth factor category;
- Calculate the mean and standard deviation from the growth factors in a given category; and
- Apply the same approach as that for seasonal factor groups to estimate precision level.

2.9 Summary

Several approaches that have been developed to incorporate seasonal effects in the calculation of total traffic volumes on a given roadway segment have been described. For models developed based on artificial intelligence technologies such as neural networks and genetic algorithms, it may be difficult to interpret resulted seasonal patterns, especially when they do not agree with engineering judgment. Although ANNs have been shown to be effective at representing complex nonlinear relationships, it is difficult to determine the relationships between variables. It is also possible to over-train a network, resulting in a memorization of the training data rather than a generalization of the relationship. Consequently, it has been recommended to use a large database for training purpose and to use proper judgment on when to stop training [SMI97]. However, such requirements generally are difficult to meet since installing and maintaining a large number of PTC sites is unlikely. Moreover, the process of determining seasonal groups cannot be automated if human judgment cannot be formulated and included in the process.

The theoretical backgrounds for the nonparametric hierarchical clustering methods described in Section 2.2.1 are relatively easy to understand. These models have been generally implemented in the practice for grouping TTMSs via popular commercial statistical software packages such as

SAS. The parametric model-based clustering models described in Section 2.2.3 are less familiar to transportation professionals. Although model-based clustering methods require more knowledge of statistics, they allow parameters measured in different scales, such as the geographical locations of the TTMSs, to be simultaneously considered in the grouping process without additional transformation. The computation engine for performing model-based clustering analysis has recently become available to the general public. For these reasons, the nonparametric hierarchical clustering and hierarchical model-based clustering methods were selected to group TTMSs for estimating MSFs in this study.

Seasonal variations in traffic are results of patterns in human activities, which are commonly influenced by land use patterns. The land use and travel behavior aspects of seasonal factors have not been adequately studied in the existing literature. By appropriately considering and incorporating roadways' functional classifications, land use, and other relevant factors into data collection and processing, it is possible to reduce the data collection effort while improving the accuracy of SF estimations.

3. TRAFFIC COUNT DATA

This chapter provides a general background of the traffic data used in this study. American Association of State Highway and Transportation Officials (AASHTO) recommends five types of traffic count data to be collected, edited, summarized, and reported during a year. They are coverage counts, long-term pavement performance (LTPP) counts, project-related counts, special count request, and data obsolescence counts [AAS92]. Following the AASHTO guidelines, FDOT maintains one of the most comprehensive traffic count programs in the country. The FDOT Transportation Statistics Office releases traffic data collected at every traffic count site on the State Highway Systems (SHS) on a Florida Traffic Information (FTI) CD-ROM. The CD provides access to nine traffic reports:

- AADT Report
- Historical AADT Report
- AADT Forecast
- 200 Highest Hour Report
- Hourly Continuous Count Report
- Annual Vehicle Classification Report
- Weekly Axle Factor Category Report
- Peak Season Factor Category Report
- Volume Factor Category Summary Report

Traffic data collected from 68 counties in Florida, including hourly traffic counts at each permanent count station, are stored in four Microsoft (MS) Access files on the FTI CD. Three out of the four MS-Access files are named in the format of Traffic_XX_YY.mdb, each containing the traffic data from those count stations in counties whose numbers fall into the range between XX and YY. Table 3 shows the database tables and the corresponding attributes stored in a Traffic_XX_YY.mdb file.

Table 3. Fields in Traffic_XX_YY.mdb

Table Name	Attributes
county	COUNTY, NAME, DISTRICT
HISTAADT	CO_SITE, COUNTY, SITE, YEAR, PTADTADJ, ASCDIR, ASCADTADJ, DSCDIR, DSCADTADJ, CLKFACT, CLPKDIRF, CLTBPCT, AADTFLG
TMSCNT	COUNTY, SITE, BEGDATE, DIR, HR1, HR2, HR3, HR4, HR5, HR6, HR7, HR8, HR9, HR10, HR11, HR12, HR13, HR14, HR15, HR16, HR17, HR18, HR19, HR20, HR21, HR22, HR23, HR24, TOTVOL, TYPE
TMSDESC	COUNTY, SITE, SECTION, LOCATION, FUNCL, SITETYPE, COMM

For this study, the last MS-Access file, i.e., Traffic_CD.mdb, contains the most critical data. Table 4 shows the data attributes contained in the Traffic_CD.mdb file. The monthly adjustment factors and day-of-week adjustment factors in the DIRECTIONAL_VOLUME table were utilized in this study. Monthly factors include JANV, FEBV, MARV, APRV, MAYV, JUNV, JULV, AUGV, SEPV, OCTV, NOVV, and DECV, each for one of the 12 months. Day-of-week factors, on the other hand, include SUNV, MONV, TUEV, WEDV, THUV, FRIV, and SATV, each for one day in a week. These data represent, in vehicle axle counts, the ratios of the

accumulated monthly totals and day of week totals to the corresponding AADT at permanent count stations. At a permanent traffic count site, AADT is the total volume of traffic on a highway segment for one year, divided by the number of days in the year. AADTs for either of the two directions as well as combined two-way volumes are recorded in the DIRECTIONAL_VOLUME table. To reveal the "truth-in-data," the data were used directly without any further adjustment or imputation whenever possible.

Table 4. Fields in Traffic_CD.mdb

Table Name	Attributes
ALL_SITES_AADT	COUNTY, SITE, YEAR, ASCDIR, ASCAADT, DSCDIR, DSCAADT, SITETYPE
ANNUAL_VEHICLE_ CLASSIFICATION	YEAR, COUNTY, SITE, CLASS, DESCRIPTION, PERCENTAGE
AXLE_ADJ_CAT	AFCAT, DESCR
COUNTY	COUNTY, NAME, DISTRICT
DIRECTIONAL_ VOLUME	COUNTY, SITE, YEAR, DIR, SUNV, MONV, TUEV, WEDV, THUV, FRIV, SATV, JANV, FEBV, MARV, APRV, MAYV, JUNV, JULV, AUGV, SEPV, OCTV, NOVV, DECV
FUNCLASS_CODE	FUNCLASS, DESCR
HI200_STATS	COUNTY, SITE, RANK, HRKFACT, HRDFACT, HIDIR, LODIR, HIVOL, LOVOL, TOVOL, HOURNO, BEGDATE
PEAKSEASON	YEAR, SFCAT, WEEK_NUMBER, DATES, VALUE, PEEK_WEEKS
PKSEAS_MOCF	SFCAT, MOCF
SEASONAL_ADJ_CAT	SFCAT, DESCR
SITETYPE_CODE	SITETYPE, DESCRIPTION
STMS_STATIONS	SFCAT, YEAR, VTTMSNO, EXCLUDE, COUNTY, SITE
SVFCAT	SFCAT, YEAR, SUNV, MONV, TUEV, WEDV, THUV, FRIV, SATV, JANV, FEBV, MARV, APRV, MAYV, JUNV, JULV, AUGV, SEPV, OCTV, NOVV, DECV, KFCTR, DFCTR, K100FCTR
TMSDESC	COUNTY, SITE, SECTION, LOCATION, FUNCL, SITETYPE, COMM
TMS_SUMMARY	COUNTY, SITE, YEAR, AADT, AADTFLG, KFCTR, KFLG, DFCTR, DFLG, TFCTR, TFLG, TRKPCT, HVYTRKPCT, MEDTRKPCT, DHTRK, DHMEDTRK, DHHVYTRK, VALID_DAYS, K100FCTR
VALID_DATA	COUNTY, SITE, HOURS, DAYS, WEEKS, MONTHS
VEEKLY_AXLE_NEW	YEAR, AFCAT, WEEK_NUMBER, DATES, VALUE, COUNTY

4. EVALUATION OF CLUSTERING METHODS

This chapter describes the processes of and findings from employing nonparametric agglomerative hierarchical clustering methods and parametric model-based clustering methods to group the TTMSs located on Florida's urban and rural roads.

4.1 Nonparametric Agglomerative Hierarchical Clustering Methods

The performance of agglomerative hierarchical clustering methods for determining seasonal factor groups is evaluated in the section. In the following subsections, the hierarchical clustering methods available in SAS are first introduced, followed by the study data and procedure used in the evaluation. The results from the hierarchical cluster analysis are then discussed. Finally the performances on the methods evaluated are summarized.

4.1.1 Clustering Methods in SAS

In current practices, seasonal factor groups are usually determined by a conventional hierarchical cluster analysis according to various similarities measures. These methods merge TTMSs into groups according to their similarities. A variety of similarity measures are used in cluster analysis. The following agglomerative hierarchical clustering methods are available in SAS Version 8 for quantifying the distance (or dissimilarity) between two clusters [SAS99]:

1. Average Linkage (AVE)
2. Centroid Method (CEN)
3. EML
4. Flexible-beta Method (FLE)
5. McQuitty's Similarity Analysis (MCQ)
6. Median Method (MED)
7. Single Linkage (SIN)
8. Ward's Minimum-Variance Method (WAR)

Each of the above clustering methods utilizes a different formula to estimate the distance between two clusters and tends to create clusters of certain types. For example, average linkage tends to join clusters with small variances and is slightly biased toward producing clusters with the same variance. Ward's method tends to join clusters with a small number of observations and is strongly biased toward producing clusters with roughly equal number of observations. The EML method is similar to Ward's minimum-variance method but is somewhat biased toward unequal-sized clusters based on practical experience. In SAS, the penalty option is used to adjust the degree of bias toward unequal-sized clusters for the EML method. The value specified as the penalty should be greater than zero. In this study, four additional penalty values other than the default value, which is 2.00, were applied in the cluster analysis to test the parameter's sensitivity to the results. These penalty values were 1.00, 1.25, 1.50, and 1.75. Additionally, two values, -0.25 and -0.50, were specified for the beta option for the flexible-beta method. A total of 13 methods were thus tested to evaluate the agglomerative hierarchical clustering methods.

In the following subsections, the process of and findings from the cluster analyses of the MSFs from the 21 TTMSs in District 4 of the FDOT, Florida with the 13 aforementioned clustering methods are described. The pseudo *F* (PSF) statistic was used to determine the number of clusters in the data. The resulted clusters from each method employed were then examined to validate their performance. The method with the optimal performance was then employed to investigate the historical clustering patterns.

4.1.2 Study Data

The traffic data from the TTMSs located within the FDOT District 4 (covering Broward, Indian River, Martin, Palm Beach, and St. Lucy counties) were investigated. The data source was the 1997-2000 FDOT Traffic Count Information CDs published by FDOT. Each CD contains the MSFs for 1997, 1998, 1999, and 2000. For 1999 and 2000, detailed hourly traffic count data are recorded on the CD. For 1997 and 1998, however, only monthly seasonal factors are available. The AADTs for these TTMSs in the four-year period vary from 2,593 to 228,518. Table 5 shows the number of available TTMSs in the study area in different years. Over the four-year period, only 19 stations consistently recorded MSF information. By including two more TTMSs located on the Florida Turnpike in the district, the MSFs from a total of 21 TTMSs were analyzed.

Table 5. Number of TTMSs in FDOT District 4 from 1997 to 2000

Year	1997	1998	1999	2000
Number of TTMSs*	29	31	27	29

* Excluding TTMSs located on the Florida Turnpike section (District 8)

4.1.3 Evaluation Procedure

The process entailed the following steps to evaluate agglomerative hierarchical clustering methods:

1. Verify seasonal factors at each TTMS;
2. Perform preliminary cluster analyses to identify outliers;
3. Perform cluster analyses and evaluate the factor groups using the data without outliers; and
4. Select the optimal clustering method to verify if seasonal groups are temporally stable.

The following subsections explain each of the steps in detail.

4.1.3.1 *Data Verification*

Before applying cluster analysis to determine seasonal groups, the historical MSFs recorded on the FDOT Traffic Count Information CDs were examined. The purpose was to identify possible outliers in the dataset by examining the temporal patterns in the data from the four consecutive years at the same TTMS. Although the days with missing data had already been excluded in the calculation of MADT, extremely low daily volumes in a given month that were most likely caused by equipment failures and other unknown reasons had not been eliminated from the data.

Consequently, the MADTs were higher than expected. Higher MADTs could also have been the result of excluding days with low traffic volumes, e.g., on weekends.

Figure 4 illustrates problems with the temporal stability of the multi-year MSFs at one of the TTMSs, Station No. 860214. The figure shows a susceptible high MSF value in September 2000. The CV[1] (coefficient of variation) for the MSFs at the station for September was 8.296% while the same statistics for the other months ranged from 0.558% to 2.663%. Figure 5 illustrates the daily volumes in September in 1999 and 2000. The volumes from the two different years were aligned together by the day of the week. As suggested by Figure 5, it is evident that the MSF in year 2000 is likely to be overestimated due to missing data since the higher volumes that tended to occur near the end of the month were excluded from the calculation of MADT, resulting in a lower than expected MSF.

To identify probable data outliers, the monthly CV for the multi-year MSFs at each TTMS was calculated first. The daily volumes were then examined for those months when their CVs were greater than or equal to 3%. The 3% threshold value was selected by observing the resulted monthly CVs from the 21 TTMSs. Since daily volumes were available only for 1999 and 2000, the MSFs in these two years may be verified. Potential outliers identified were replaced with the median MSF of the respective month that was determined from the multi-year data by assuming the MSFs in the 1997 and 1998 datasets being accurate.

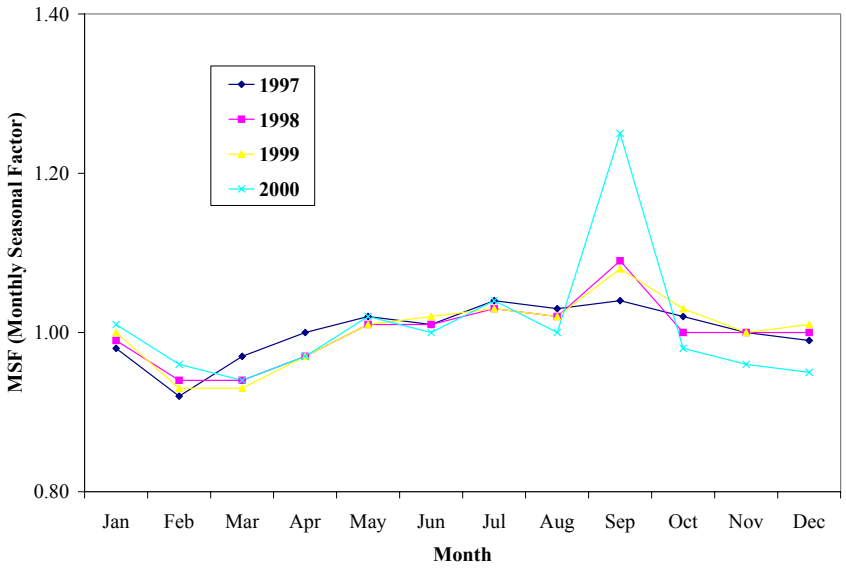


Figure 4. MSF vs. Month at Station 820614 in 2000

[1] Defined as the standard deviation divided by the mean MSF and multiplied by 100 to get a percentage.

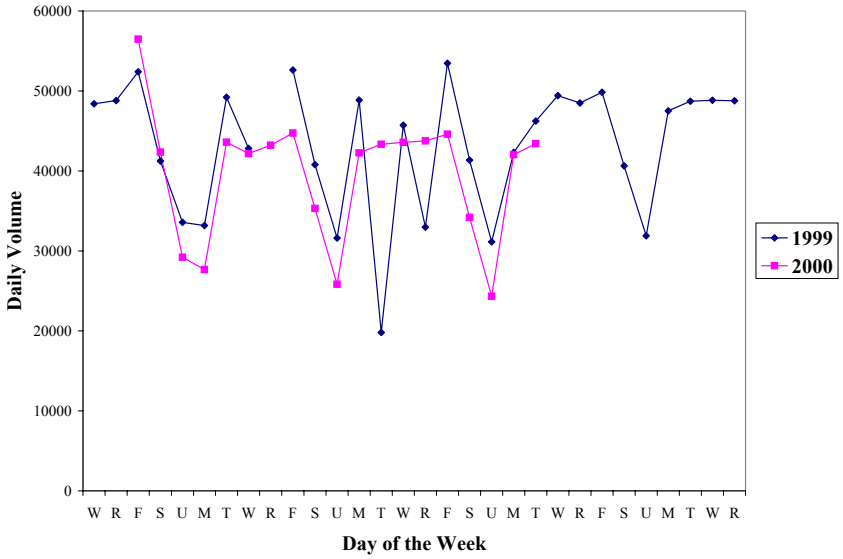


Figure 5. Daily Volume versus Day of the Week[2] at Station 820614 in 1999 and 2000

4.1.3.2 *Preliminary Cluster Analysis*

After the MSFs were verified, the year 2000 MSFs were analyzed with the various clustering methods to identify outliers. The pseudo F (PSF) statistic was used in the study as the criterion to determine the number of clusters in the data. The relatively large PSFs, i.e., a local peak in the graph of the PSFs plotted against the number of clusters, indicate a stopping point. The PSF for a given level is calculated as follows [SAS99]:

$$PSF = \frac{\dfrac{T - P_G}{G - 1}}{\dfrac{P_G}{n - G}} \tag{43}$$

where

T = $\sum_{i=1}^{n} \left\| x_i - \bar{x} \right\|^2$;

P_G = $\sum W_j$, where summation is over the G clusters at the G^{th} level of the hierarchy;

G = number of clusters at a given level of the hierarchy;

n = number of observations (i.e., TTMSs);

x_i = i^{th} obervation;

$\bar{x}$ = sample mean vector;

W_k = $\sum_{i \in C_k} \left\| x_i - \bar{x}_k \right\|^2$; and

$\bar{x}_k$ = mean vector for cluster C_k .

[2] M for Monday, T for Tuesday, W for Wednesday, R for Thursday, F for Friday, S for Saturday, and U for Sunday.

T is the sum of squared Euclidean distances from each observation to the overall mean, while P_G is the Euclidean distance measured from the observations in a given cluster to its cluster mean [JOH02]. The resulted clusters were examined after the preliminary cluster analyses. The TTMSs belonging to a single member cluster were treated as outliers and excluded from the evaluation process since their monthly fluctuation patterns were significantly different from the others.

4.1.3.3 *Evaluation*

The 13 clustering methods were applied again to the year 2000 data after the TTMS outliers were eliminated. The resulted clusters from each method were evaluated based on the pooled estimate of common variance, which was simply the arithmetic average of the monthly MSF variances for the TTMSs in the same group. The spatial locations of the TTMSs clustered in the same seasonal factor groups by the method with the least pooled variance were examined to verify if they were logical and reasonable. The method(s) that yielded reasonable results was then used to cluster the MSFs for the data from other years.

4.1.3.4 *Temporal Stability*

Outliers were first identified and excluded by performing a preliminary cluster analysis using the method(s) that derived the least pooled variance from the year 2000 data. The remaining TTMSs were than analyzed again to verify if cluster groups were stable from year to year.

4.1.4 Results and Discussions

All the clustering methods revealed that Stations 860306, 890259, and 940144 were outliers during the preliminary analysis step. As a result, these stations were excluded in the subsequent analysis. Additionally, methods such as AVE, CEN, and SIN were more robust to outliers than the other hierarchical methods. The results suggested that the AVE, CEN, and SIN methods might be preferable for screening out outliers. The 13 clustering methods were then applied again to analyze the data from a total of 18 TTMSs. Table 6 presents the resulted PSFs obtained after excluding the outliers.

In Table 6, the cells with relatively large PSFs, which suggested a possible stopping point from merging groups further, are highlighted. Table 6 indicates that different penalty values did not alter the results from the EML method, since they resulted in the same PSFs. Moreover, regardless of what beta value was specified, the same stopping point was obtained by the FLE method. To confirm that the same PSFs represented the same clustering groups, the resulting tree structure diagrams that indicated the disjoint clusters at a specified level from the EML and FLE methods were examined. They were found to be identical. Therefore, the default value in SAS may adequately serve the purpose of constructing seasonal factor groups. Table 6 also shows that almost all methods indicated that the optimal number of clusters was five. From practical experience, it is also unlikely to have fewer than two or more than seven seasonal factor categories [TMG01]. For these reasons, the pooled variances were calculated only for those clusters that were numbered between two and seven even if the PSF criterion suggested more than one optimal number of clusters for a method such as the MED method.

Table 6. PSFs at Different Hierarchical Levels for Various Clustering Methods

Number of Cluster	AVE	CEN	EML					FLE		MCQ	MED	SIN	WAR
			1	1.25	1.5	1.75	2	-0.25	-0.5				
17	10.90	10.90	10.90	10.90	10.90	10.90	10.90	10.90	10.90	10.90	10.90	10.90	10.90
16	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
15	9.20	9.20	9.20	9.20	9.20	9.20	9.20	9.20	9.20	9.20	9.20	9.20	9.20
14	8.60	8.20	8.60	8.60	8.60	8.60	8.60	8.60	8.60	8.60	8.20	8.20	8.60
13	8.30	8.30	8.40	8.40	8.40	8.40	8.40	8.40	8.40	8.30	8.30	7.20	8.40
12	8.50	8.50	8.50	8.50	8.50	8.50	8.50	8.50	8.50	8.50	8.50	4.70	8.50
11	8.80	8.80	8.80	8.80	8.80	8.80	8.80	8.80	8.80	8.80	8.80	5.40	8.80
10	9.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00	8.70	9.00	6.10	9.00
9	9.20	9.00	9.20	9.20	9.20	9.20	9.20	9.20	9.20	9.20	9.00	4.30	9.20
8	9.60	8.40	9.60	9.60	9.60	9.60	9.60	9.60	9.60	9.60	7.70	5.10	9.60
7	10.30	7.80	10.30	10.30	10.30	10.30	10.30	10.30	10.30	10.30	8.20	5.90	10.30
6	10.50	8.40	10.50	10.50	10.50	10.50	10.50	10.50	10.50	9.40	5.60	6.30	10.50
5	10.70	8.40	10.70	10.70	10.70	10.70	10.70	10.70	10.70	9.70	6.90	3.90	10.70
4	8.90	10.90	9.60	9.60	9.60	9.60	9.60	8.90	9.60	8.90	7.40	4.50	9.60
3	6.60	6.60	9.30	9.30	9.30	9.30	9.30	7.70	9.30	6.40	8.60	6.60	9.30
2	5.40	5.40	10.20	10.20	10.20	10.20	10.20	7.20	10.20	7.20	8.20	5.40	10.20

In order to compare the resulting clusters from different methods, the pooled variances were also computed for the CEN, MED, and SIN methods for which five was not the optimal number of clusters. For simplicity, the 13 clustering methods were categorized into five method groups according to their clustering results as follows: Group 1 – AVE, EML, FLE, and WAR; Group 2 – CEN; Group 3 – MCQ; Group 4 – MED; and Group 5 – SIN. Table 7 shows the calculated pooled variance for each group at different levels of the hierarchy. The shaded cells indicate the pooled variance for the number of clusters recommended by the PSF criterion. The cells shaded with horizontal lines indicate more outliers, i.e., single TTMS in a cluster, which were detected and eliminated from the calculation of pooled variance. The "crossed-out" cells were thus discarded due to the difference in sample size. Table 7 indicates that the Group 3 method, i.e., the MCQ method, produced the least pooled variance at five clusters and was consequently defined as the optimal clustering method.

Table 7. Pooled Variance for Various Clustering Groups

Method Group	Number of Clusters			
	3	4	5	6
1			4.37	
2		4.92	4.78	
3			4.18	
4	7.4		4.22	
5	6.09		5.08	4.02

The spatial dispersion patterns of the TTMS clusters from the MCQ method were then examined. Figure 6 illustrates the seasonal factor groups determined by the MCQ method for five cluster

groups. Each TTMS is labeled with a number between 1 and 5 to indicate its seasonal factor category.

As illustrated in Figure 6, the two TTMSs on the Florida Turnpike were assigned to Category 1. Category 2 included the TTMSs that were located on the major roads close to Florida Turnpike but on the west side of it. Category 3 included the two TTMSs that were located on the major roads far west in Palm Beach County. The two TTMSs located on the major roads near the boundary between Palm Beach and Martin counties were assigned to Category 4. The last category covered nearly all of the rest of the TTMSs, including four located on the Interstate 95 and six on major roads that were on the east side of Florida Turnpike and/or I-95, whichever were further west. The spatially clustering patterns shown in Figure 6 suggested that it was not appropriate to merge the TTMSs on the Florida Turnpike section with those located on the regular major roads or interstate highways. Additionally, roadway functionality did not seem to play an important role in determining seasonal groups. It was the spatial location of a given TTMS that matters since a TTMS tended to be clustered with those in its proximity. Since the seasonal groups determined by the MCQ method were logical and reasonable, the method was thus implemented to analyze the MSFs collected from 1997 through 1999.

The three TTMSs that were previously excluded from the year 200 data were identified as outliers for the data from 1997 through 1999. The three stations were thus excluded from the three-year dataset. Figures 7 to 9 illustrate the seasonal factor groups determined by the MCQ method for years 1997, 1998, and 1999, respectively, after the outlier stations were eliminated. The resulted numbers of seasonal groups were seven, six, and six for Figures 7, 8, and 9, respectively.

Figures 7 to 9 show that the MCQ method did not consistently assign a TTMS to the same group from year to year. However, the change in the grouping patterns during the four-year period revealed a gradual shift in spatial clustering patterns from north-south direction to east-west direction. For example, although Figure 7 does not show any significant spatially clustered pattern among the TTMSs in 1997, Figure 8 shows that two relatively larger groups, i.e., Groups 5 and 6, were formed at the north and south sides of the district. Figure 9 shows that the spatially clustering patterns began to shift from the north-south direction to the east-west direction, similar to what was illustrated in Figure 6 from the year 2000 data. One of the probable reasons for such a shift could be because the traffic first began to circulate in the major activity centers within each county in 1997. The traffic interchanging between different counties became increasingly significant with time, which resulted in increased utilization of the interstate freeways, including I-95 and Florida Turnpike. Similar seasonal fluctuation patterns were thus observed for the TTMS along the freeways. After the freeways became congested around 1998 and 1999, traffic began to shift to local streets. Since the land use and developments along the east and west sides of the I-95/Florida Turnpike were different, two distinguished seasonal patterns were thus formed.

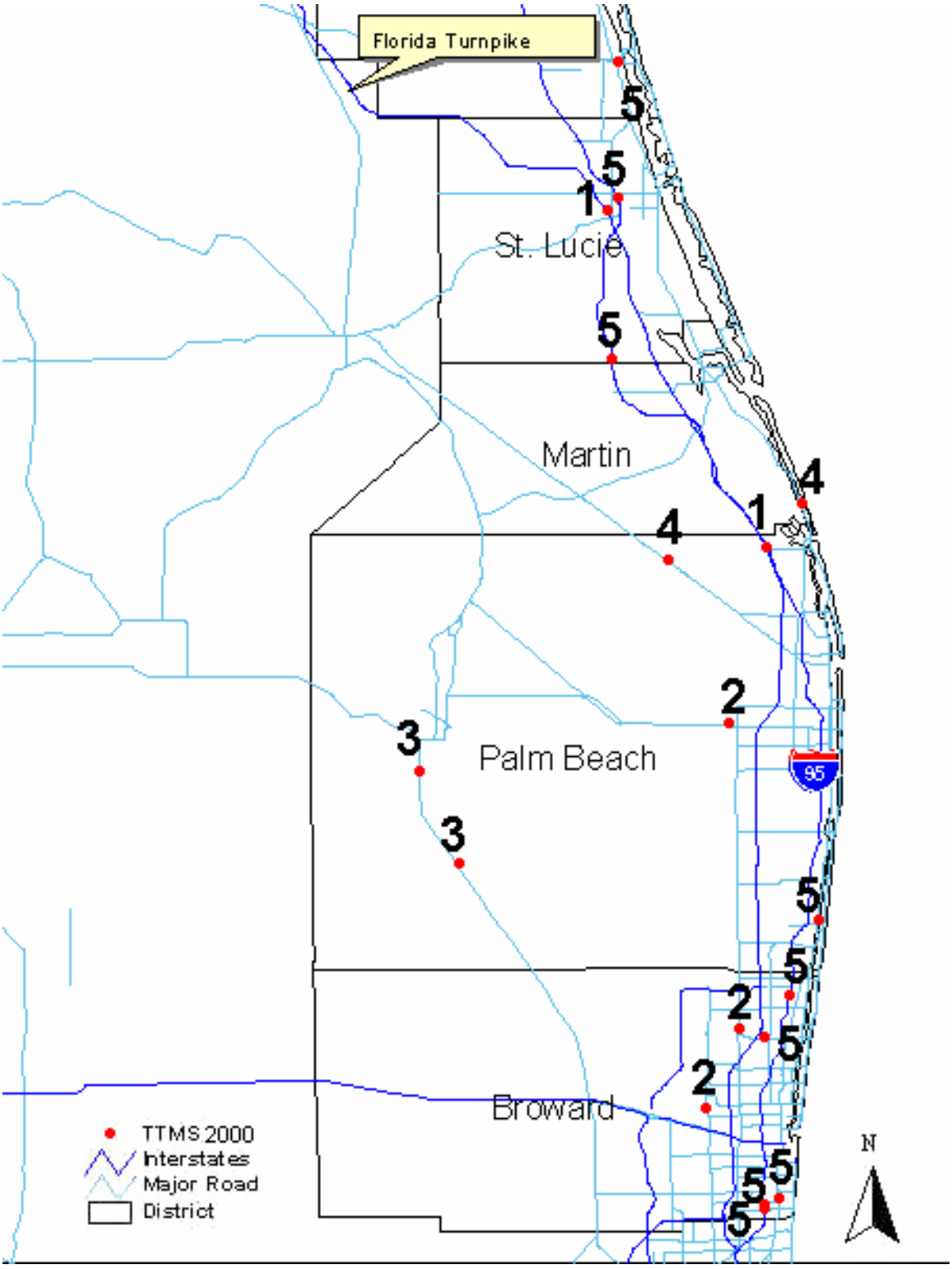


Figure 6. Seasonal Cluster Groups Determined by the MCQ Method (Year 2000)

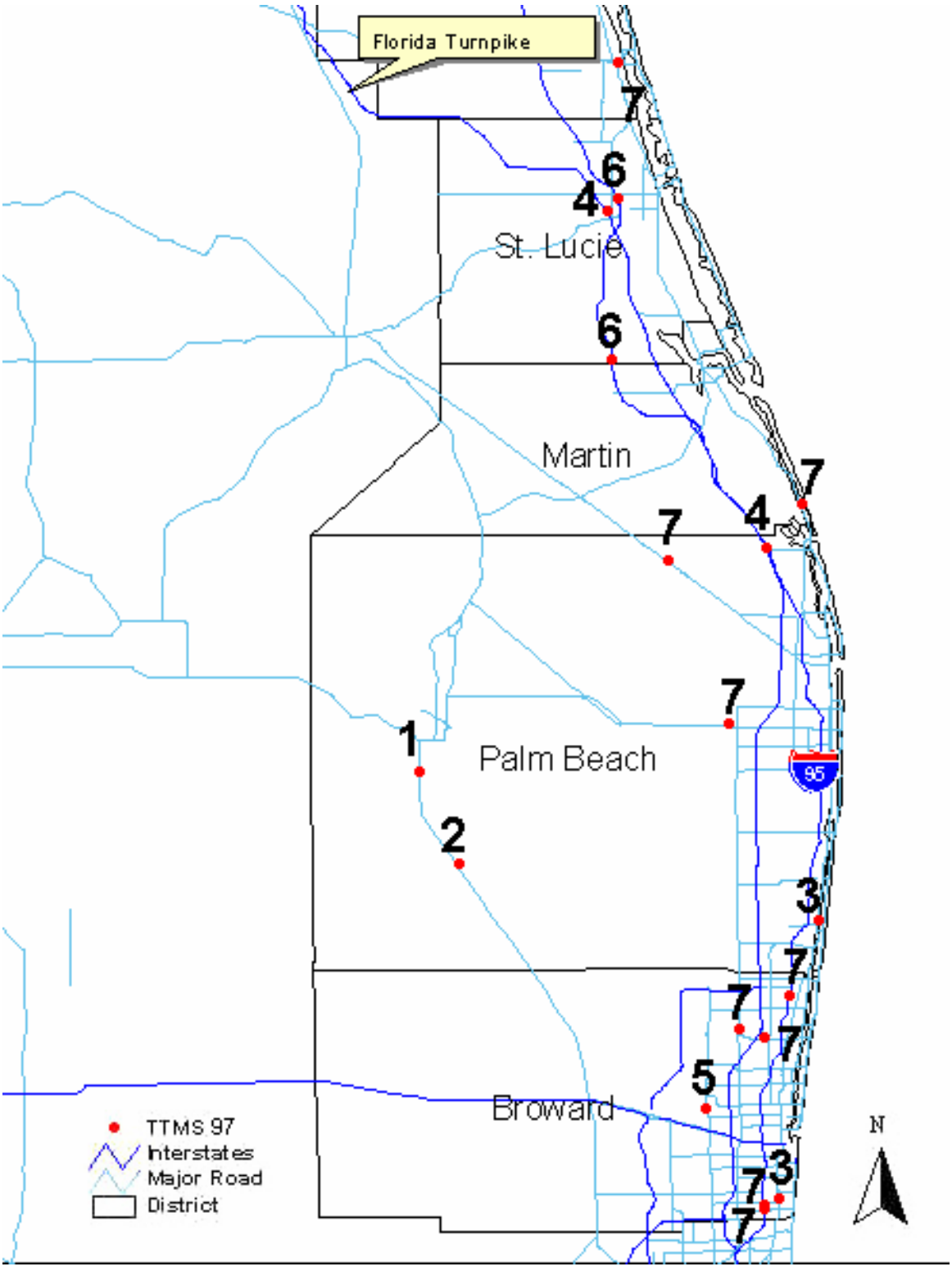


Figure 7 Seasonal Cluster Groups Determined by the MCQ Method (Year 1997).

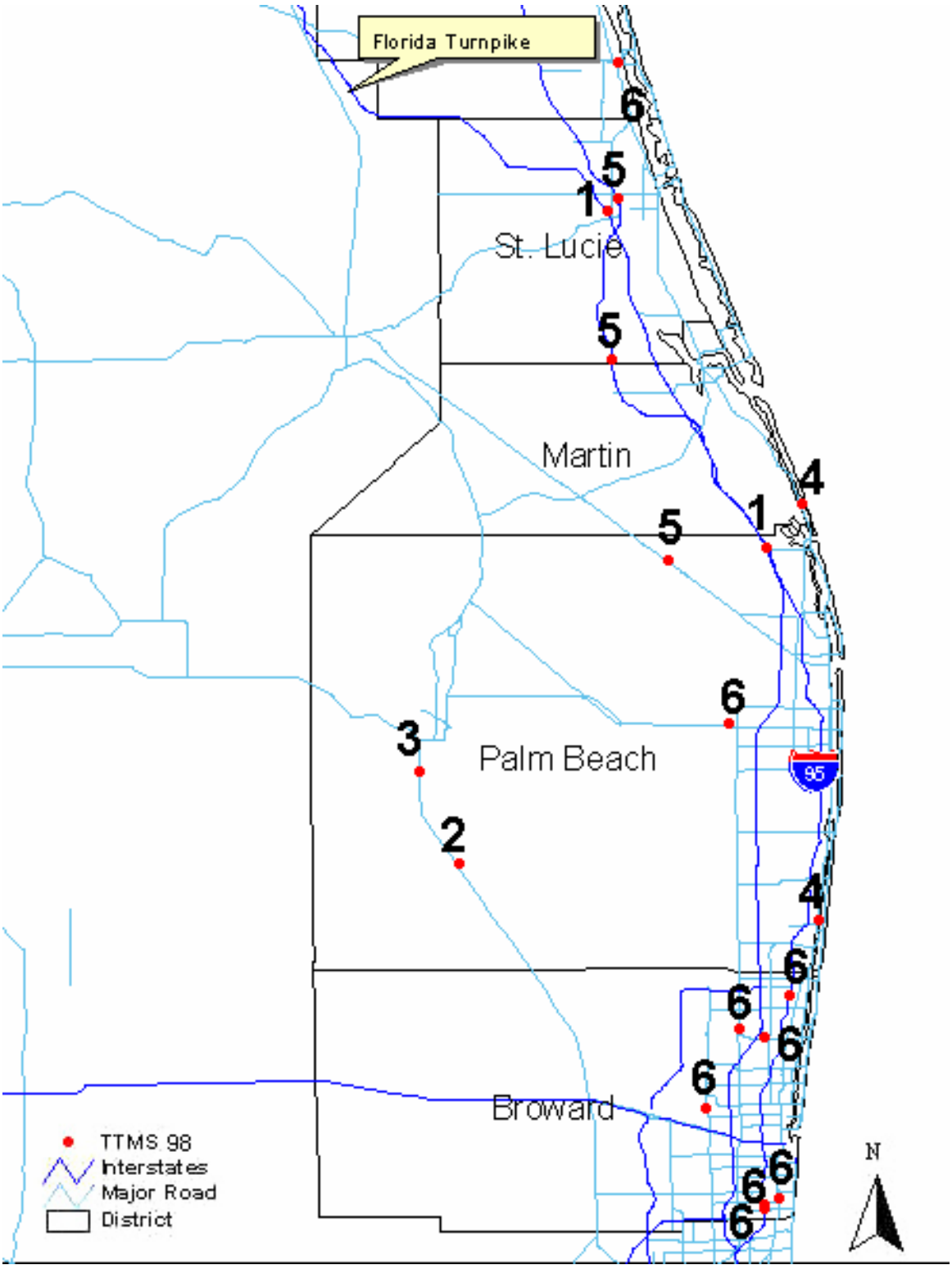


Figure 8 Seasonal Cluster Groups Determined by the MCQ Method (Year 1998).

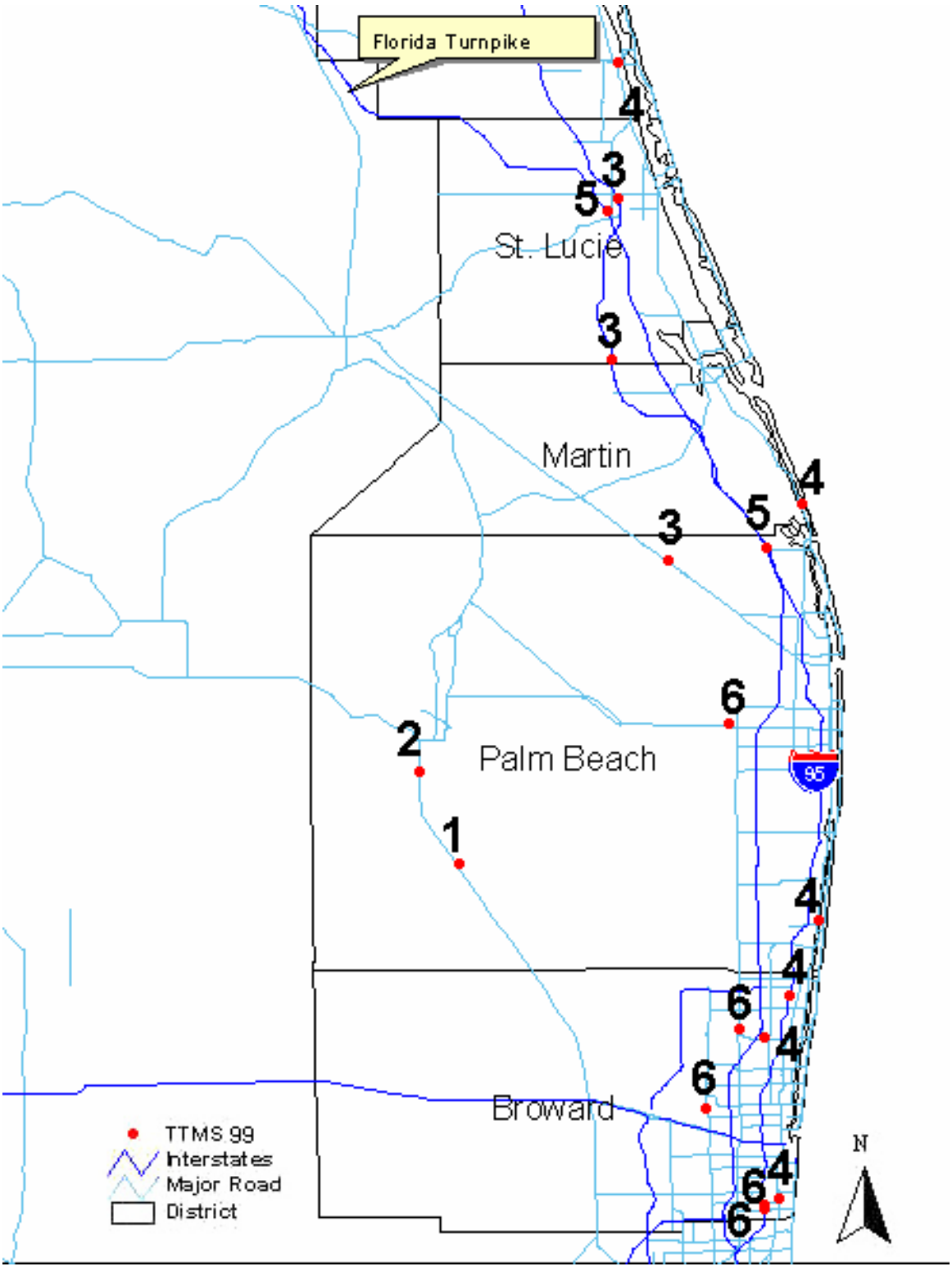


Figure 9 Seasonal Cluster Groups Determined by the MCQ Method (Year 1999).

4.1.5 Summary

A total of eight agglomerative clustering methods were evaluated. The average linkage, centroid, and single linkage methods were found to be more robust to outliers than the other methods. The study also found that the McQuitty's (MCQ) method performed better than the other methods on grouping TTMSs after outliers were eliminated. Although the results from analyzing the four-year MSF data with the MCQ method showed that the compositions of seasonal groups were not stable over time, the change in the spatially clustering pattern indicated that more variables should be included in the process of determining seasonal cluster groups. The study also found that roadway functionality did not seem to play an important role in determining seasonal groups. It was the spatial location of a given TTMS that mattered since TTMSs tended to be clustered with those in its proximity. Based on data from FDOT District 4, the PSF statistic was found to be a good measure to determine the number of clusters after potential outliers were excluded. Engineering judgment is still necessary in determining the grouping via hierarchical clustering analysis. These methods, however, provide a starting point for the FDOT staff to fine tune the seasonal factor groups.

4.2 Parametric Model-Based Hierarchical Clustering Methods

This section presents the process and findings from a model-based clustering analysis of the 12 MSFs, each corresponding to a given month in a year, from 129 permanent count stations on rural roads in Florida. In the following sections, the model-based cluster methods available in MCLUST, an extension of the SAS software, are first presented. The study data are then briefly described, followed by the analysis procedures adopted in this study [FRA02]. The results from the model-based clustering are discussed. Finally, a summary is provided.

4.2.1 Clustering Methods in MCLUST

The model-based clustering was accomplished using the MCLUST software developed by the University of Washington [FRA02]. A total of ten models, each specified with a unique set of geometric features of the covariance matrix as described in Table 2, were implemented to analyze the 129×12 MSF matrix. They were EII, VII, EEI, VEI, EVI, VVI, EEE, VVV, EEV, and VEV models. The definitions of these models are described in Section 2.2.3. As previously mentioned, the covariance matrix for the parameters in a given cluster may be decomposed into matrices that determine the orientation, volume, and shape of a distribution. The models incorporated in MCLUST allowed the characteristics of distribution to vary between clusters or maintain the same for all clusters. The MCLUST software was designed as an extension to the S-Plus program in either a Windows or UNIX environment. The S-Plus program must be installed to allow the execution of the MCLUST module.

4.2.2 Study Data

In this study, the 2002 traffic data from the TTMSs located in Florida rural areas were analyzed. Again, the data were obtained from the Traffic Count Information CD distributed by the FDOT. After excluding those with incomplete data, the MSFs from a total of 129 TTMSs were retrieved

and stored in a 12-element matrix, one element for each month (i.e., 129×12). The geographical coordinates were also retrieved for each count station from the 2002 FDOT Traffic CD.

4.2.3 Evaluation Procedure

The analysis process entailed the following steps to evaluate model-based clustering:

- Apply model-based strategy for clustering 2 to 100 groups using the models listed in Table 2; and
- Add the coordinates of each TTMS in the data matrix and perform the model-based clustering again.

First, the model-based strategy adopted in MCLUST for clustering was used. Three explicit procedures [FRA02] were employed in model-based clustering strategy in this study. The first procedure involved the application of model-based agglomerative hierarchical clustering to approximately maximize the unconstrained classification likelihood (also known as the VVV model) to create the initial count site classifications for the expectation-maximization (EM) algorithm. The second procedure implemented the Expectation-Maximization (EM) algorithm beginning with the classification from hierarchical agglomeration since reasonably good partitions were commonly produced from the first procedure and no other information about the groupings was available. The Bayesian Information Criterion (BIC) was then calculated. These three procedures were iteratively executed for numbers of seasonal factor groups ranging from two to 100 and the resulted BIC statistics were used to determine the optimal number of clusters.

In the second step, another data matrix (129×14) with the geographical location of each permanent count station in terms of X and Y coordinates as the additional characteristics was analyzed using the same procedure. The purpose was to investigate the incorporation of the spatial locations of permanent count stations in the grouping analysis. The clusters obtained from each method were then examined to evaluate its performance with the assistance of GIS.

4.2.4 Results and Discussions

Figures 10 and 11 illustrate the BIC values for different numbers of groups ranging from 2 to 100 for the 129×12 and 129×14 data matrices, respectively, for the various models. The BIC values from the VVV model were not included because the model produced no factor group for the 12- (12 MSFs) matrix and only one factor group for the 14-component (12 MSFs plus two coordinates) matrix. The figures show that models including VII, VEI, EVI, VVI, EEV, VEV, and VVV did not provide feasible solutions for every number of groups specified. The reason was that non-positive definite covariance matrices were encountered in the EM iterations. A matrix is positive definite if all of its eigenvalues are positive. For symmetric matrices such as covariance matrices, positive definiteness is assured if the matrix and every principal sub-matrix have a positive determinant. As described earlier, the EM algorithm breaks down when the covariance matrix corresponding to one or more seasonal factor groups is singular or near singular. Since there may exist two or more permanent count stations with nearly the same MSFs, linear dependency was unavoidable. However, both figures reveal that feasible solutions were obtained for every number of groups from the EII, EEI, and EEE models.

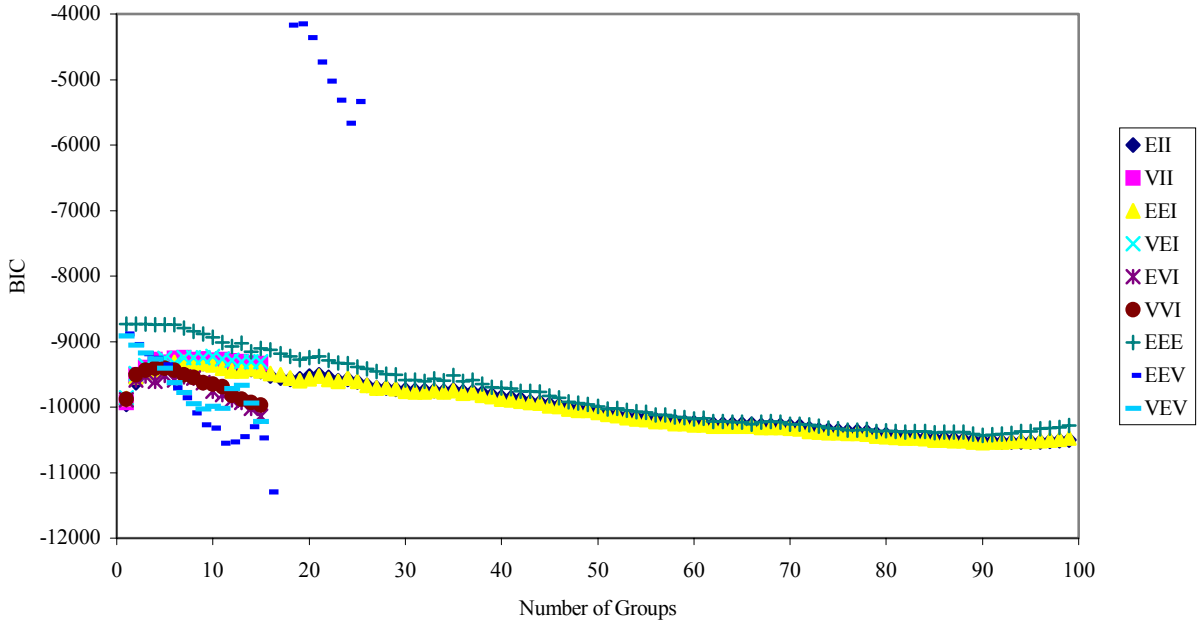


Figure 10. BICs for 12-Component Data

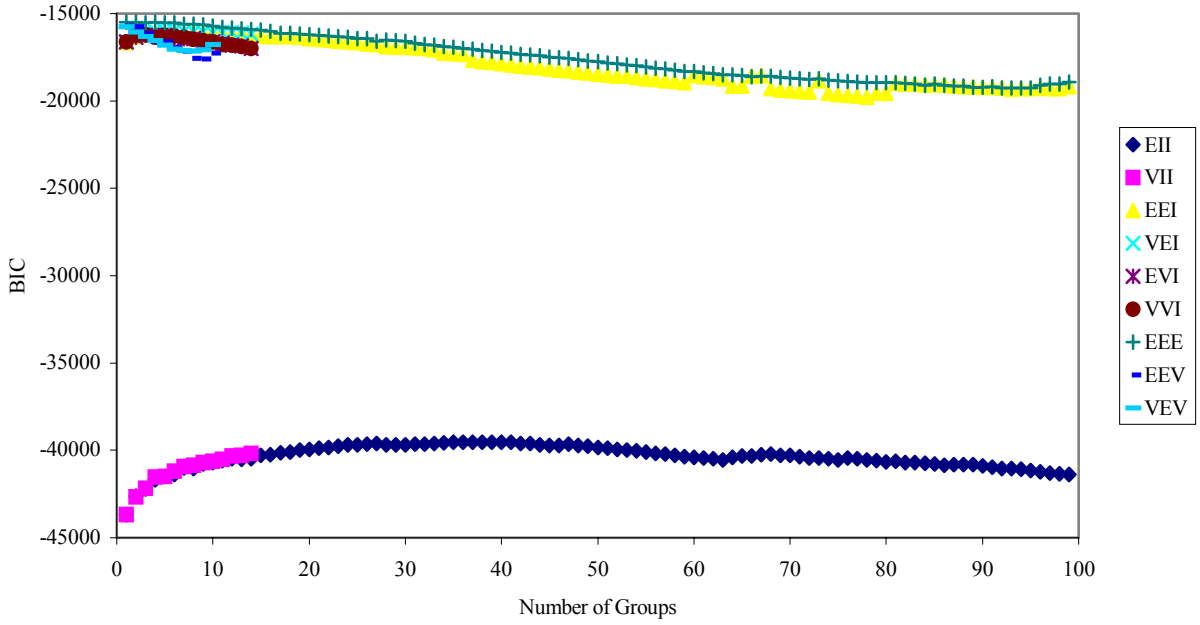


Figure 11. BICs for 14-Component Data

Table 8 provides the optimal number of groups and the associated BIC values for the various models. The tolerance for relative convergence of the likelihood was 10^{-6}. The original default value for the tolerance was 10^{-5} in the MCLUST software. A smaller tolerance value was applied here in an attempt to derive better results. Table 8 shows that the EEV and VVV models produced the best BIC values for the 12- and 14-component matrices, respectively. Additionally,

more seasonal factor groups were obtained from the 14-component matrices from the EII and VII models.

Table 8. Optimal Numbers of Groups and BICs for Various Models (tolerance = 10^{-6})

Comp		EII	VII	EEI	VEI	EVI	VVI	EEE	EEV	VEV	VVV
12	Group	9	8	8	11	6	6	3	20	2	NA
	BIC	-9313	-9249	-9313	-9226	-9517	-9408	-8733	-4149	-8911	—
14	Group	38	15	9	9	5	4	2	2	2	2
	BIC	-39518	-40194	-16070	-16006	-16233	-16196	-15486	-15708	-15746	-14946

To further explore the performance of model-based clustering for the ten models, the mean MSF and the thresholds defined as ±10% of the mean MSF of each cluster were calculated. The threshold is currently applied by the FDOT to determine if TTMSs are satisfactorily classified. As shown in Table 9, the total numbers of cases for TTMSs classified into a given group with their MSFs exceeding the upper and lower thresholds for a given month were calculated for the optimal number of groups. The fist measure of effectiveness (MOE) was the cumulative number of months that exceeded the thresholds. The second MOE was the total number of TTMSs from the months that exceeded the thresholds. The results given in Table 9 showed that five of the 240 (i.e., 20×12) months exceeded the thresholds for the EEV model for the 12-component matrix. It appeared that the EEV model was superior to the other models when only the 12 MSF data from the rural roads in Florida were used in the analysis. For the 14-component matrix, the EII model produced groupings with a negligible misclassification error. Although there tended to be more misclassifications in terms of frequencies when the locations were considered in the grouping process, the magnitude varied among different datasets and models since the optimal numbers of groups also generally increased.

As previously mentioned, the mixed model that simultaneously considered MSFs and coordinates of locations estimated the probability that a TTMS belonged to the k^{th} seasonal factor group. Consequently, the final membership of a TTMS in a group was based on the maximum probability. To visualize the spatial distribution of the groups, a geographical centroid was determined for each factor group. A circle was then drawn on a map to show the area that was approximately occupied by a given seasonal factor group. The radius was the distance between the centroid and the TTMS in the group that was farthest from the centroid. Figure 12 illustrates the three MSF groups from the 12-component matrix for the EEV model, which produced the optimal BIC value among the 10 models investigated. The results showed that the EEV model would produce clusters that included TTMSs located more than 400 miles apart. It is unlikely in practice that TTMSs located so far apart would be grouped together. Therefore, merely using the MSF data with the model-based approach was not sufficient to determine the seasonal factor groups.

Table 9. Frequencies and Percentages of Possible Misclassifications

Comp	Possible Misclassification		EII	VII	EEI	VEI	EVI	VVI	EEE	EEV	VEV	VVV
12	MOE 1[1]	Count	10	15	9	14	19	19	26	5	18	
		%	9.26%	15.63%	9.38%	10.61%	26.39%	26.39%	72.22%	2.08%	75.00%	—
	MOE 2[2]	Count	15	19	29	23	47	77	103	8	176	
		%	0.97%	1.23%	1.87%	1.49%	3.04%	4.97%	6.65%	0.52%	11.37%	—
14	MOE 1	Count	30	43	11	17	24	20	16	16	14	14
		%	6.58%	23.89%	10.19%	15.74%	40.00%	29.17%	66.67%	66.67%	58.33%	58.33%
	MOE 2	Count	56	76	36	35	99	107	198	208	192	184
		%	3.62%	4.91%	2.33%	2.26%	6.40%	6.91%	12.79%	13.44%	12.40%	11.89%

1. MOE 1: Cumulated number of months with observations exceed the corresponding 10% thresholds of average MSFs
2. MOE 2: Cumulated number of TTMS with MSFs exceed the 10% thresholds of average MSFs

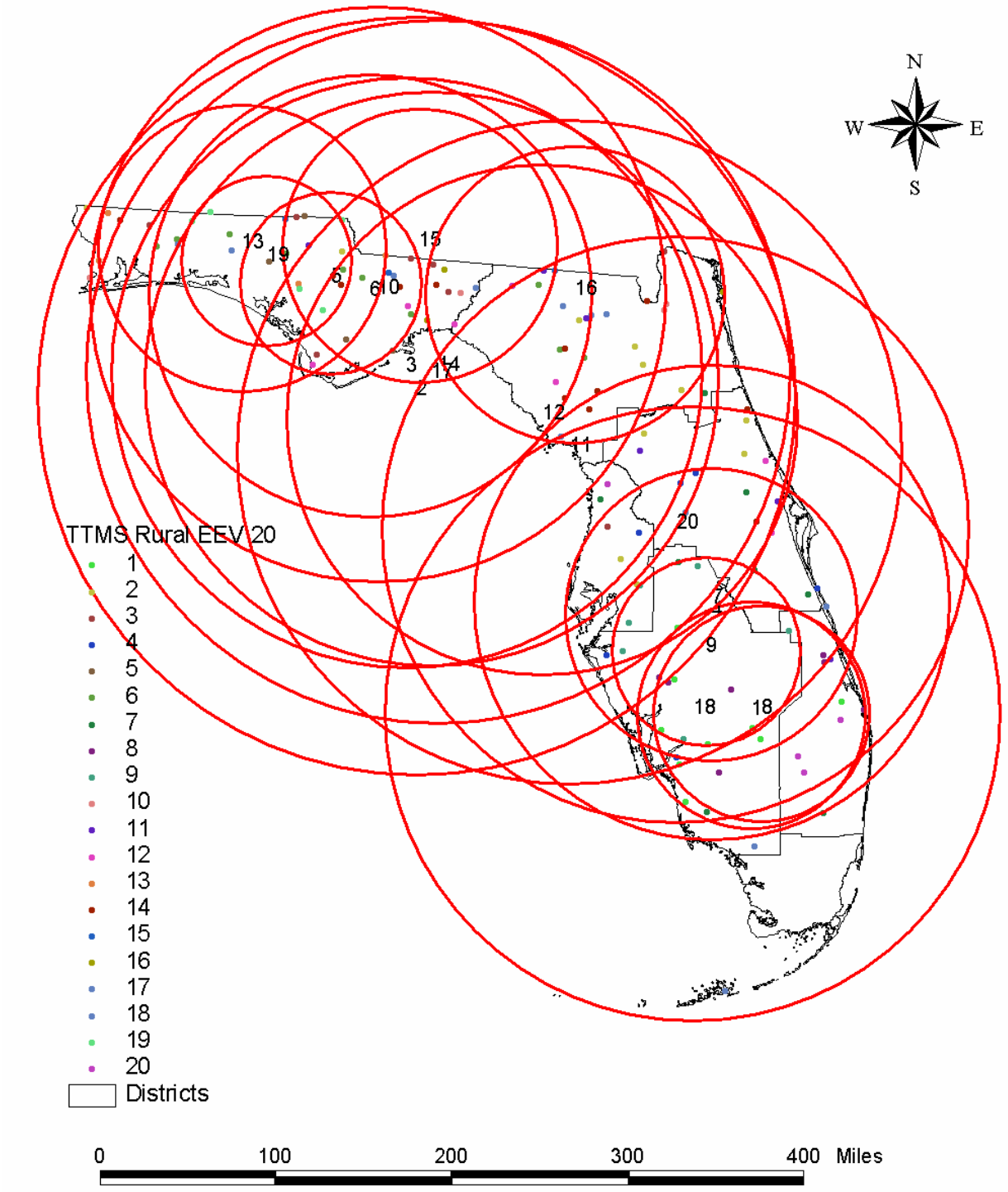


Figure 12. Twenty-Group EEV Classifications from 12-Component Matrix

Figure 13 shows a more spatially clustered pattern of the TTMSs from the VVV model when the coordinates of the TTMSs were included in the analysis. Although the VVI model produced the optimal BIC for the 14-component matrix, it was impractical to use only two seasonal factor groups (see Table 9) because high variations in MSFs were introduced. The EII and VII models, which produced more groups, were subsequently investigated. Figures 14 and 15 show the grouping results for the 14-component matrix from the EII and VII models, respectively. Since spatial coordinates were incorporated to determine the grouping, TTMSs tended to be clustered with those in their proximity. Moreover, in comparison with the VII model, the radii of the groups from the EII model were relatively consistent. This was because the equal volume of the covariance matrix was assumed in the EII model. Based on the results presented in Table 9 and Figures 12 through 15, it may be concluded that the EII model is the best model to implement for grouping the TTMSs in the rural areas of Florida. The TTMSs that were classified into the same groups had relatively similar seasonal fluctuation patterns and were also located close to each other.

4.2.5 Summary

The results from evaluating the performance of model-based clustering methods for seasonal factor grouping showed that, without additional information such as the spatial locations of the TTMSs, the model-based clustering methods such as the EEV model could produce classifications with a negligible grouping error of 2.08% when statewide MSF data were used in the analysis. However, the TTMSs in the same group were also scattered spatially, in some cases over 400 miles apart. By incorporating coordinates of the TTMSs in the model-based clustering, it was found that the EII and VII models produced practical numbers of factor groups. By further comparing the MSFs in the factor groups derived from these two models, the EII model was identified as the one with the best performance since it produced the least grouping error.

The results from the systematic analysis of the model-based clustering described here may be considered as a reasonable starting point for determining the seasonal factor groups in practice. The procedure provides practitioners with greater flexibility in classifying a TTMS since the probability for a TTMS belonging to a given factor group is estimated. For example, if a TTMS is considered misclassified, it may be easily reassigned to the next factor group according to the sequence determined by the grouping probability. Moreover, incorporating the coordinates of TTMSs in the model-based cluster analysis allows geographical effects to be considered in the grouping process and groups of TTMSs to be derived that not only share similar MSF patterns but are also located close to each other. The results will benefit transportation professionals in assigning a seasonal factor group (category) to a short count site by considering spatial proximity. The model-based clustering process presented in this study will also allow other characteristics such as land use that could not be considered in the conventional grouping approaches to be incorporated into the grouping process.

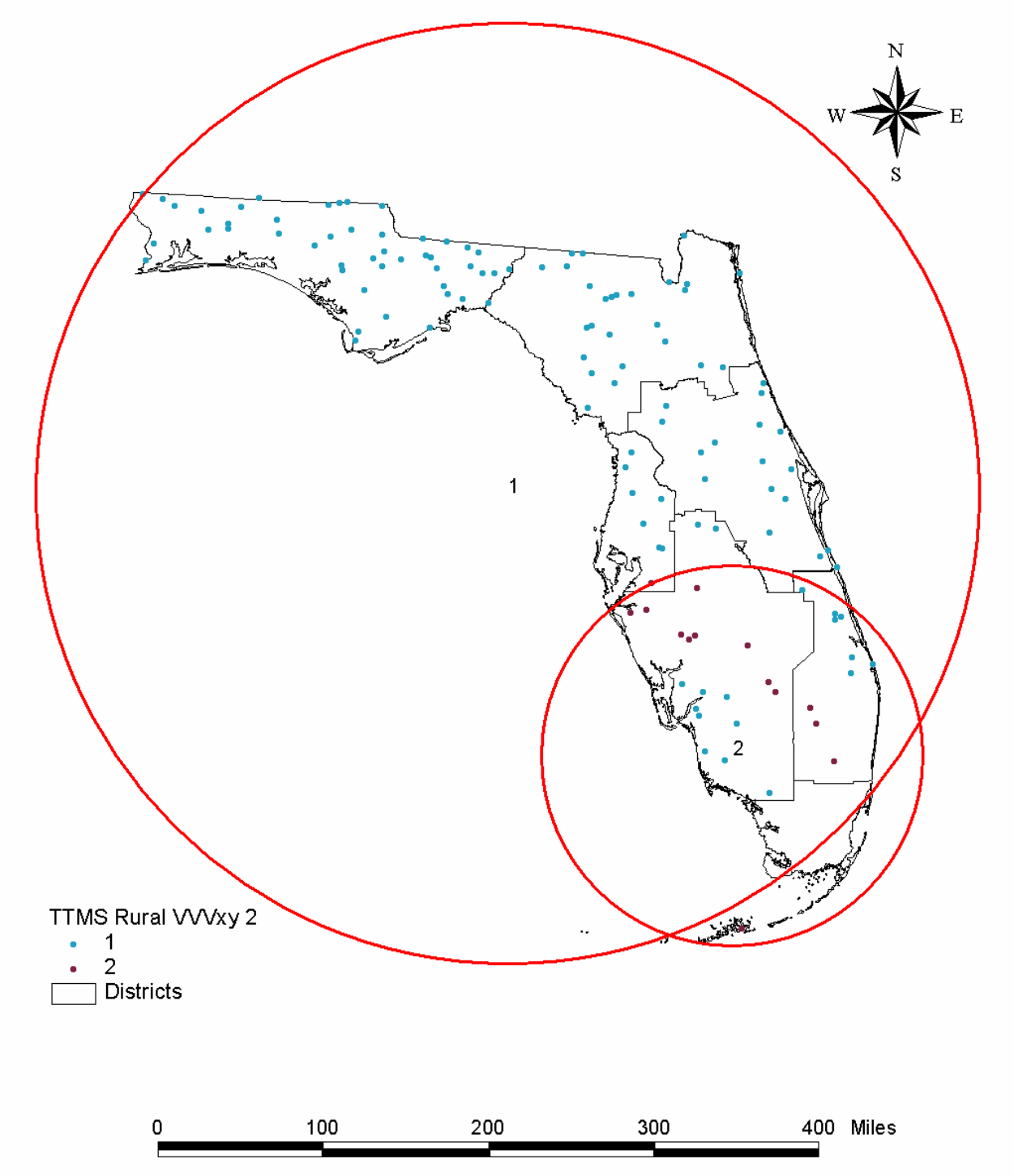


Figure 13. Two-Group VVV Classifications from 14-Component Matrix

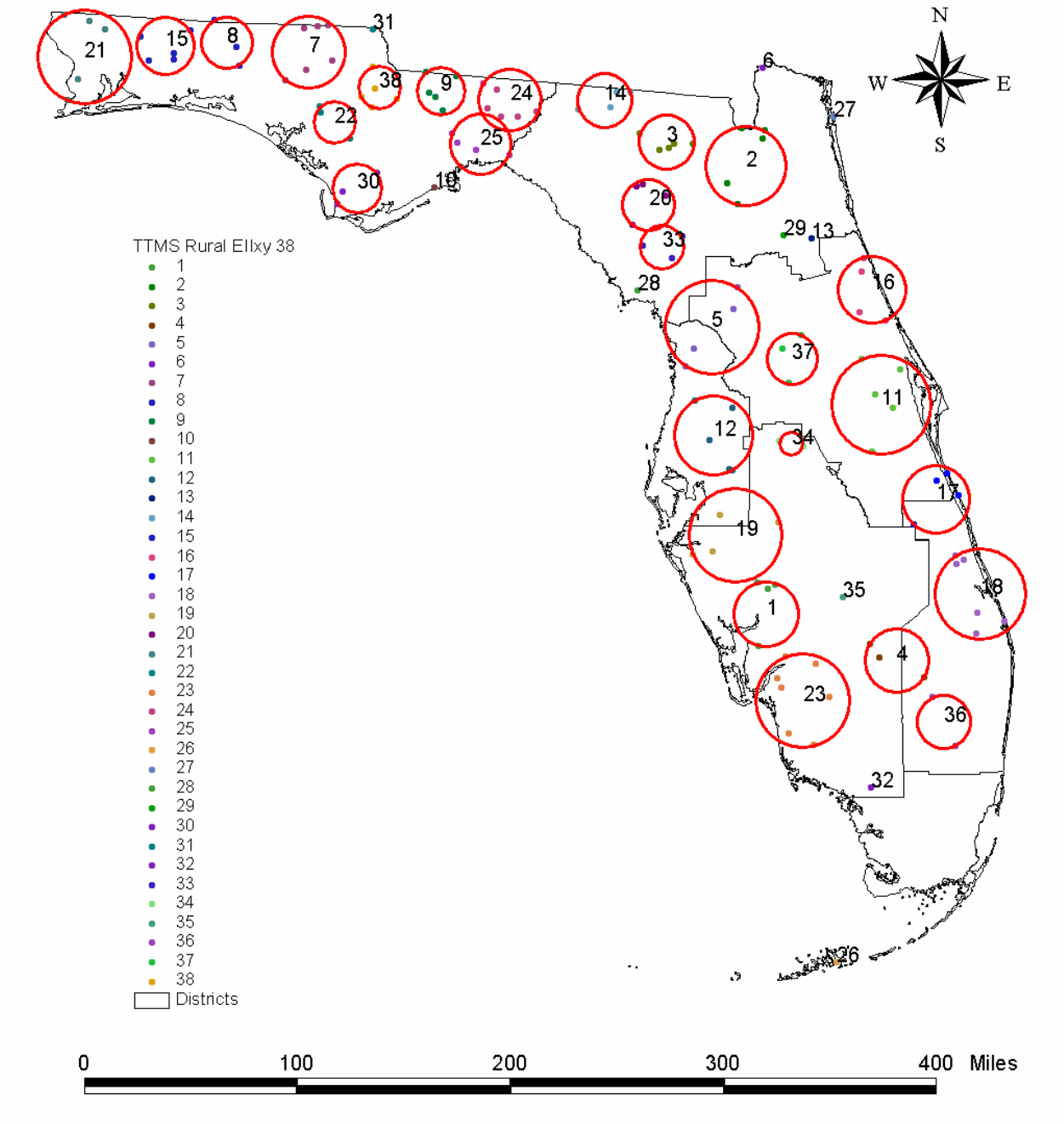


Figure 14. Thirty-Eight -Group EII Classifications from 14-Component Matrix

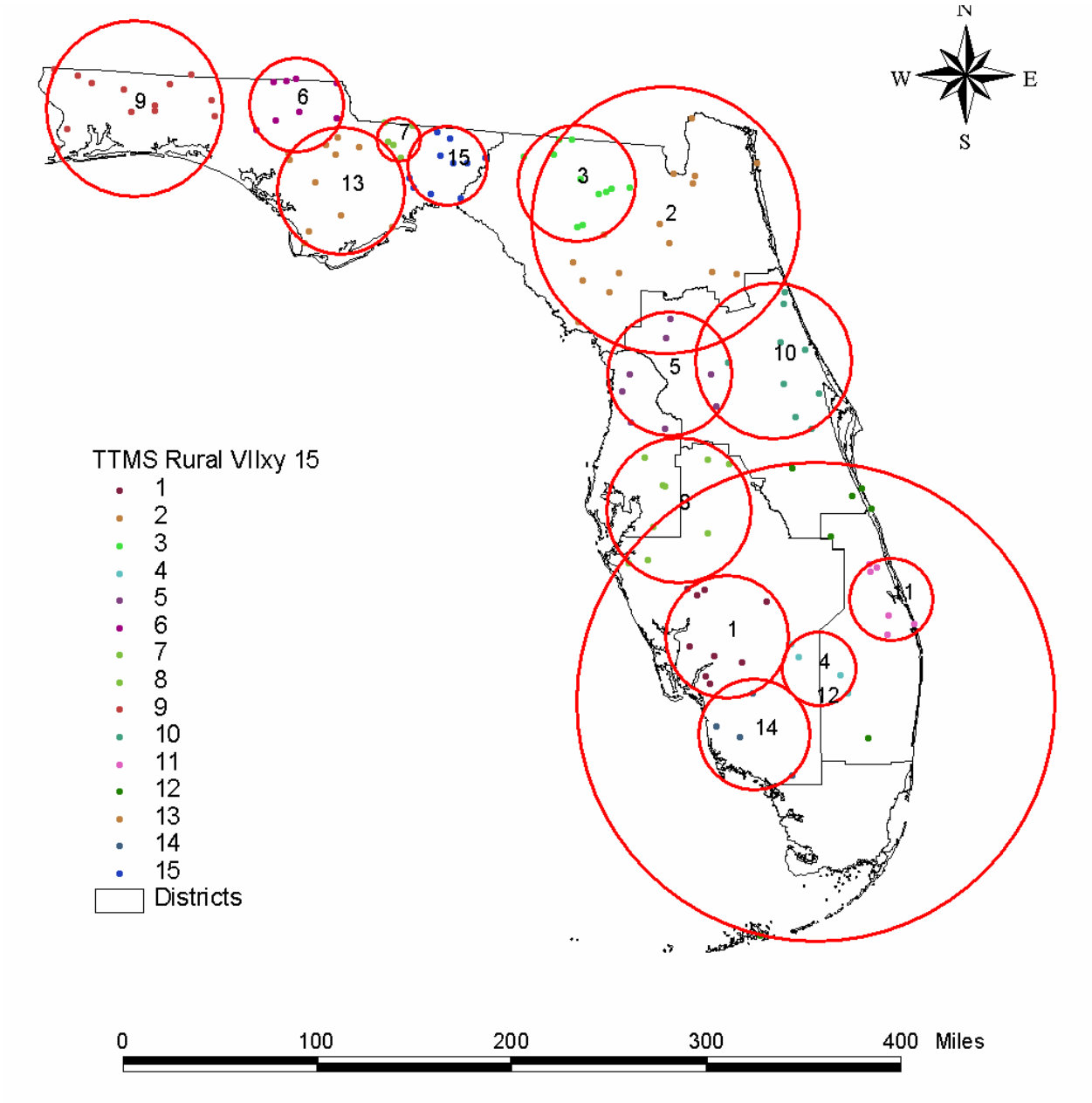


Figure 15. Fifteen-Group VII Classifications from 14-Component Matrix

5. SEASONAL FACTOR GROUP ASSIGNMENT

In this chapter, factors that may be helpful in identifying the seasonal factor group for a short-term traffic count site are described. The factors considered in this study may be classified into two major groups: land use and geographical location. In Section 5.1, the land use factors that have well-known effects on travel patterns are first identified via conventional multiple linear regression analysis on the MSFs collected at a given month at TTMSs located on selected urban roads in Florida. The parameters identified as significant factors may be used to determine the seasonal factor category to which a short count site belongs, assuming that these factors have similar effects on the seasonal variability and traffic characteristics at the short-term and permanent count sites. Section 5.2 describes the process and findings from using the same analysis approach described in Section 5.1 on the MSFs collected at the TTMSs on selected rural roads in Florida. Section 5.3 presents a fuzzy decision tree method for assigning seasonal factor categories to short-count sites.

5.1 Urban Area Regression Analysis

5.1.1 Introduction

This section presents the results from applying multiple linear regression analysis on the MSFs for a given month that were collected from the TTMSs on selected urban roads in Florida. The urban roads were defined as those in an urban area in Florida. The "urban" shape file from the 2002 Traffic Information CD was used to identify the TTMSs that were located on the urban roads. A total of 71 urban areas were included in the GIS theme file, as shown in Table 10.

Table 10. Florida Urban Areas

Arcadia	Fort Myers	Lehigh Acres	Pahokee	Stuart
Avon Park	Fort Pierce	Live Oak	Palatka	Sun City Center
Belle Glade	Fort Walton Beach	Marathon	Panama City	Tallahassee
Beverly Hills	Gainesville	Marco	Pensacola	Tampa
Bonita Springs	Homosassa Springs	Mariana	Perry	Tavares
Brooksville	Immokalee	Melbourne	Plant City	Titusville
Clermont	Inverness	Miami	Punta Gorda	Vero Beach
Clewiston	Jacksonville	Middleburg	Quincy	West Palm Beach
Crestview	Key Largo	Mount Dora	Ruskin	Winter Haven
Dade City	Key West	Naples	Sarasota	Yulee
Daytona Beach	Lady Lake	North Port	Sebastian	Zephyrhills
Defuniak Springs	Lake City	Ocala	Sebring	
Deland	Lake Wales	Orange City	Spring Hill	
Deltona	Lakeland	Orlando	St. Augustine	
Fernandina Beach	Leesburg	Ormond Beach	Starke	

As previously stated, SF groups are currently assigned to short-term traffic count sites based on a site's geographical location and its functional roadway classification in Florida. While spatial proximity is possibly a reasonable assumption, there have not been adequate studies to determine if it is the only factor in determining seasonal groups. A more objective and data-oriented

approach needs to be developed in order to explain the underlying causes of seasonal fluctuation patterns in traffic data and to allow a sound process of assigning seasonal factors to short period counts based on factors in addition to spatial proximity. In the following sections, the development of multiple linear regression models for estimating seasonal factors is described. The method has the potential to reduce the subjective nature in the current practice by either allowing seasonal factors to be estimated directly for each short-count station or helping assign seasonal groups to short counts.

There is little documented evidence in the literature regarding land use variables as predictors for seasonal traffic patterns. Although some studies have pointed to trip characteristics as important to seasonal traffic patterns, it is unclear how such characteristics are related to the basic patterns of land use and demographic and socioeconomic characteristics. Since traffic is the result of human activities, however, land use and demographic and socioeconomic characteristics are likely to be influential in the nature of the activities, thus travel and traffic patterns. One of the objectives of this study was to investigate which land use, demographic, and socioeconomic characteristics associated with TTMS are important in determining MSFs. The quantification of the impact of these characteristics was determined using GIS techniques and multiple linear regression analysis.

5.1.2 Study Data

The study area was the Southeast Florida tri-county urban area, which encompasses Broward, Miami-Dade, and Palm Beach counties. It had a population of five million according to the 2000 census. In a large urban area, the types of transportation facilities and the land use patterns are complex. Consequently, the traffic patterns are more varied. Some important characteristics of the area are its tourist-oriented economy and seasonal residents, who live in northern states for most of the year and come to Florida to spend the winter months. The retired population (defined as population aged 65 and over) is also important. Table 11 illustrates the demographic characteristics of the tri-county area population and households (HHs) based on the 2000 census. Although the retired population in the entire tri-county area did not seem to be particularly large in terms of percentage, their total number was close to one million. If their spatial distribution was not random but clustered, the difference in their trip making could possibly produce an impact on the seasonal variations locally.

Table 11. Tri-County Demographics in 2000

	Population	Population over 65 (%)	HHs	Seasonal HHs (%)*
Miami-Dade	2,253,362	13.33	777,378	4.03
Broward	1,623,018	16.09	654,787	7.62
Palm Beach	1,131,184	23.2	474,295	11.64
Tri-County	5,007,564	16.45	1,906,460	7.15
Florida	15,982,378	17.56	6,341,121	8.12

* Ratio of seasonal households to permanent households

For this study, the dependent variables were the 12-month MSFs, which were obtained from the continuous traffic count data collected at 27 TTMSs in the study area and published by FDOT in the 2000 Traffic Information CD. Among the 27 TTMSs used in the regression analysis, 10

were on freeways, 12 on principal arterials, three on minor arterials, and two on collectors. Figure 16 illustrates the distribution of the 27 TTMSs.

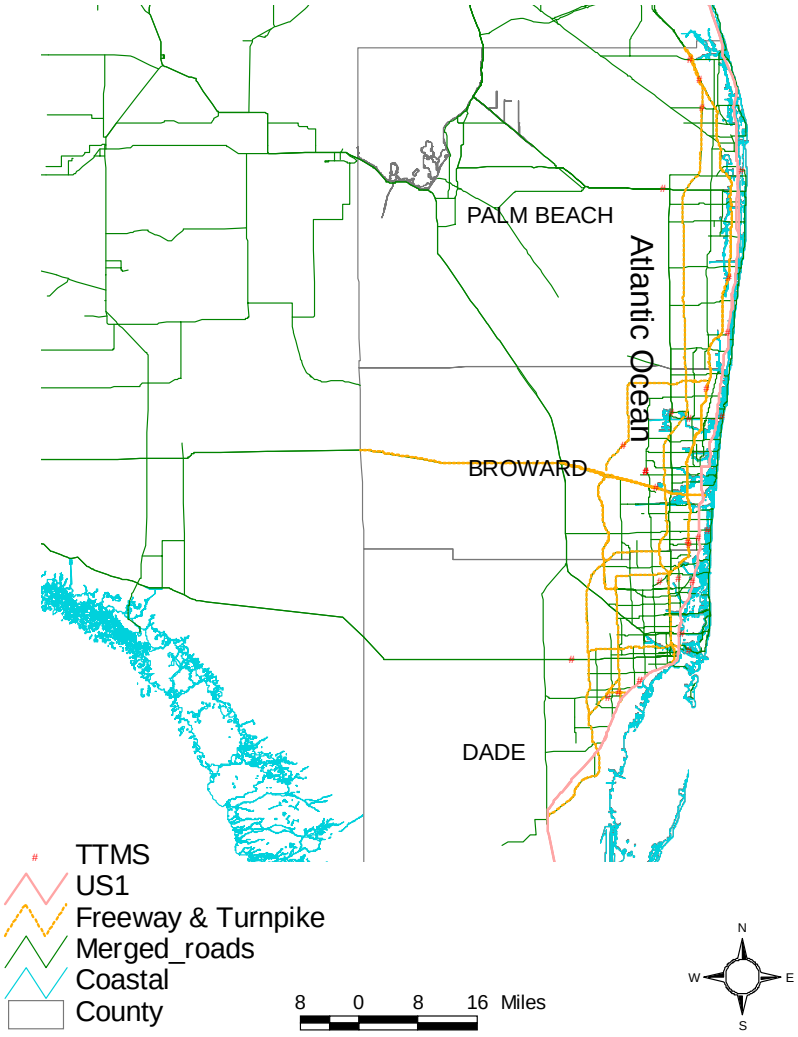


Figure 16. TTMSs in the Tri-County Urban Area

Potential independent variables used in regression analysis were those likely to have a causal relationship with seasonal factors. They described the demographic and socioeconomic characteristics of an area where a TTMS was located. They were selected based on two major considerations: (1) whether the source data were readily available or could be collected easily and economically for both base and forecast years; and (2) whether variables could be quantified. The independent variables prepared to develop the multiple regression models could be classified generally into the following categories:

- Roadway characteristics;
- Aggregate demographic and socioeconomic variables in the surrounding area of count stations;

- Geographic spatial location dummy variables from the cluster analysis.

The data used to compile these variables included the following:

- Population, number of occupied hotel/motel rooms, industrial employment, commercial employment, service employment, total employment, and school enrollment at TAZ level estimated by county planning departments for their 1999 transportation models.
- Population, number of retired householders by different income groups, number of seasonal households, number of total households, and number of total housing units from the 2000 census at census tract level.
- Employment data for the year 2000 from a commercial employment database purchased by FDOT. The data included, for each business establishment, the business name, address, location, business type (identified by a SIC code), number of employees, etc.
- Street network with federal functional classification.

The independent variables are described in the following subsections.

5.1.2.1 *Roadway Characteristic Variables*

Variables in this category are summarized in Table 12. The data were obtained from the 2000 FDOT Traffic Information CD and the Roadway Characteristics Inventory (RCI) database. Four variables, DFR, DPA, DMA, and DCO, were dummy variables that indicated the type of road where a TTMS was located.

Table 12. Roadway Characteristic Variables for Urban Roads

Variable	Description
L	Number of lanes
AADTPL	AADT per lane
DFR	Equals 0 if TTMS not located on urban freeway; 1 otherwise
DPA	Equals 0 if TTMS not located on urban principle arterial; 1 otherwise
DMA	Equals 0 if TTMS located on urban minor arterial; 1 otherwise
DCO	Equals 0 if TTMS not located on urban collector; 1 otherwise

5.1.2.2 *Demographic and Socioeconomic Variables*

It is well known that socioeconomic conditions affect the travel behavior of trip makers. The variables in this category were designed to reflect the socioeconomic characteristics of the population in areas around a count station. The use of buffer methods was based on the assumption that traffic at a count station was affected by trips generated in or attracted to the area within a certain distance of that count station. Traffic may be made up by local and regional (through) traffic, although this depends on the definition of local or immediate impact area. Buffer methods will not be able to account for the characteristics of all the traffic generators in the region, which is a limitation of the buffer methods. However, in the absence of more accurate yet simple practical methods, buffer methods appear to be a reasonable tool for this application.

The variables were compiled using three different buffer methods:

- Buffer Method 1. A circular buffer around each count station was created. The buffer radii varied according to the functional classification of the roadway segments that each count station was located [ZHA01] to reflect the service area size of different types of roads. The buffer radii were one mile for principle arterials, 0.5 mile for minor arterials, 0.25 mile for collector, which were based on the common spacing of roads of different function classes. A larger buffer zone implied that the MSFs for a count station were impacted by the characteristics of a larger surrounding area. It was difficult to determine the impact area of a freeway, and radii of one through ten miles with a one-mile increment were tested.

- Buffer Method 2. This method was similar to Buffer Method 1 but differed in the way in which buffers were created for count stations located on freeways. Instead of creating a circular buffer around a count station on a freeway, where there may or may be any access from the local roads to the freeway section, buffers were created around the intersections of freeway ramps and local streets within 10, 15, and 20 miles of the count station with the buffer radii varying according to the functional classifications of the streets connected with the ramps, as in the first buffer method.

- Buffer Method 3. Instead of circular buffers, linear buffers were created around the roadway segments where the count stations were located. The buffer width was determined based on the functional classifications of the roadway segments as in Buffer Method 1, while the length of the buffer was defined by the two intersecting streets on either side of the count station with the same or higher functional classification. For freeways, the buffers were created in the same manner as in Buffer Method 2.

Using the above three buffer methods, the following independent variables were compiled.

Percentage of retired households by different income levels

Retired households were defined as households with a retired householder and were categorized into groups based on the age of the retired householder, which were further divided into subgroups by household income level. Table 13 defines the five groups of retired households and their income levels. The Group 1 variables were percentages of households (HHs) with retired householders aged 65 and over at four income levels defined in the 2000 census. This group of variables was calculated as:

$$RHP1\_LI = \frac{\text{Number of Retired HHs of Low Income Level}}{\text{Number of Total HHs}} \tag{44}$$

$$RHP1\_MLI = \frac{\text{Number of Retired HHs of Medium Low Income Level}}{\text{Number of Total HHs}} \tag{45}$$

$$RHP1\_MHI = \frac{\text{Number of Retired HHs of Medium High Income Level}}{\text{Number of Total HHs}} \tag{46}$$

Table 13. Variables for Retired Households with Different Income Levels

Group	Variables	Age*	Income Level Definition	Notes
1	*RHP1_LI* *RHP1_MLI* *RHP1_MHI* *RHP1_HI*	≥ 65	Low Income level – $0 - $24,9999 Medium Low Income Level – $25,000 - $44,999 Medium Low Income Level – $45,000 - $99,999 Medium Low Income Level – $100,000 and over	2000 census definition
2	*RHP2_LI* *RHP2_MI* *RHP2_HI*	≥ 65	Low Income Level – 0 - $20,000 Medium Income Level – $20,000 - $40,000 High Income Level – $40,000 and over	U.S. standard definition
3	*RHP3_LI* *RHP3_MI* *RHP3_HI*	65 to 74	Same as Group 2 income level definition	U.S. standard definition
4	*RHP4_LI* *RHP4_HI*	65 to 74	*Miami-Dade County* Low income – under $35,966 High income –$35,966 and over *Broward County* Low income – under $41,691 High income –$41,691 and over *Palm Beach County* Low income – under $45,062 High income –$45,062 and over	Countywide median HH income
5	*RHP5_LQ* *RHP5_MLQ* *RHP5_MHQ* *RHP5_HQ*	≥ 65	Similar to Group 4 but with four income levels instead of two	Countywide median income quartiles

* Age of Retired Householder

$$RHP1\_HI = \frac{\text{Number of Retired HHs of High Income Level}}{\text{Number of Total HHs}} \tag{47}$$

The second group of variables was similar to the first group, but differed in the income level definition. Group 2 variables were *RHP2_LI*, *RHP2_MI*, and *RHP2_HI*, which represented the ratios of retired households of the three income levels to the total households. The third group of variables, *RHP3_LI, RHP3_MI*, and *RHP3_HI,* was similar to the Group 2 variables except that only retired households with householders aged between 65 and 75 were considered. The fourth group of variables, *RHP4_LI* and *RHP4_HI*, was the percentages of households with householders aged 65 to 75 and with low or high income level defined based on countywide median income. The Group 5 variables, *RHP5_LQ*, *RHP5_MLQ*, *RHP5_MHQ*, and *RHP5_HQ,* were, respectively, the percentages of households with retired householders aged 65 or older and with household income falling into the low, medium low, medium high, and high quartiles of median household income.

Population Density

Variable *POPD* was the population density around a count station inside its buffer zone.

Seasonal Household Percentage

Variable *SHP* represented the seasonal households as a percentage of permanent households in a buffer zone around a count station.

Hotel/Motel Rooms

HMP3 and *HMP4* reflected the number of hotel and motel rooms and hotel population, respectively, around a count station and were calculated as follows, where the coefficient of 1.61 was the average hotel room occupancy factor based on data from Miami-Dade County and was applied to all three counties in the study area:

$$HMP4 = \frac{\text{Total Occupied Hotel/Motel Rooms}}{\text{Total Occupied Hotel/Motel Rooms + Total HHs}} \tag{48}$$

$$HMP3 = \frac{\text{Total Occupied Hotel/Motel Rooms} \times 1.61}{\text{Total Occupied Hotel/Motel Rooms} \times 1.61 + \text{Total Population}} \tag{49}$$

The total households and total population in the above formulae were to deal with cases when there were no hotels/motels in a buffer zone around a count station. Moreover, including the total households and population in the denominators also allowed the relative significance of tourist population to be reflected.

School Enrollment

SED was the size of school enrollment in a buffer area around a count station, and *SEP* was school enrollment as a percentage of population plus school enrollments.

Employment Variables

Fifteen variables describing employment in a buffer area surrounding a count station are listed in Table 14.

Table 14. Employment Variable Definitions for Urban Roads

Variable	Description
INDD	Number of industrial workers per acre in a buffer zone around a count station
COMD	Number of commercial workers per acre in a buffer zone around a count station
SERD	Number of service workers per acre in a buffer zone around a count station
COMSERD	Number of commercial plus service workers per acre in a buffer zone around a count station
INDP1	Industrial workers as a percentage of total workers
INDP2	Industrial workers as a percentage of total workers plus population
COMP1	Commercial workers as a percentage of total workers
COMP2	Commercial workers as a percentage of total workers plus population
SERP1	Service workers as a percentage of total workers
SERP2	Service workers as a percentage of total workers plus population
COMSERP1	Commercial/service workers as a percentage of total workers
COMSERP2	Commercial/service workers as a percentage of total workers plus population
EMPD	Number of workers per acre in a buffer zone around a count station
RETAILP	Retail workers as a percentage of total retail workers plus population
HOTELP	Hotel workers as a percentage of total hotel workers plus population
HOTELCAMP	Hotel/RV-camp workers as a percentage of hotel/RV-camp workers plus population

5.1.2.3 *Geographic Spatial Location Dummy Variables*

From cluster analyses described in Chapter 4, it was determined that count stations that belonged to the same seasonal factor groups tended to locate in the same general geographic areas [LI03]. This suggested that if geographic boundaries could be identified to define areas with similar seasonal traffic patterns, they could be used as a basis to assign established seasonal groups to short counts. To test the strength of the location variables representing different geographic area where a count station was located, six dummy variables, *DG1* through *DG6,* were created for the six seasonal groups obtained from the cluster analyses.

5.1.3 Multiple Linear Regression Analysis

Using the three buffer methods, a total of 16 datasets were compiled. Before regression analyses were carried out, the variables were screened to eliminate those that were highly correlated. For instance, the income variable with the highest correlation with MSFs was selected and the others

were eliminated from the datasets. Stepwise regression was applied in the variable selection procedure. Significant levels were set at 0.05 for regressors to enter and stay in a model. Multicollinearity among independent variables was checked based on variance inflation factors (VIFs), which should be less than 10.

First, the location dummy variables were excluded to test the explanatory powers of the land use variables. For the different buffer sizes tested for freeways, Buffer Method 1 with a five-mile buffer size was found to yield a model with the highest adjusted R^2, while Buffer Method 2 and 3 with a buffer size of 15 miles gave the best results. Regardless of which buffer method was used to compile the socioeconomic variables, the significant variables and the adjusted R^2 did not vary significantly across the three models. For most of the months, the Buffer Method 1, which was also the simplest method among the three buffer methods, gave the highest adjusted R^2. No good models were found for May, October, and November. Moreover, no models were found for December regardless which buffer method was used. The models with the highest R^2 based on the three buffer methods are presented in Table 15. The shaded cells indicate highest adjusted R^2 among all three models for the given month.

Table 16 summarizes the variables of the three sets of models with the signs of their coefficients indicated in the parentheses. As the ratio of AADT to MADT, a higher MSF indicates a lower MADTA and vice versa. A positive sign of a coefficient indicated that an increase in the corresponding variable would result in a larger MSF thus relatively fewer trips and a negative sign a smaller MSF thus more traffic. It may be seen that the variables and their signs were mostly consistent across the three sets of models. Presence of hotels/motels in an area was a major traffic-inducing factor in the warm winter months (January through April) in southeast Florida, which was the high season for visitors and for local residents to recreate. Therefore, the coefficients for variables *HMP3* and *HMP4* are negative for these months. It is likely that out-of-town visitors began to increase in October, but the effects of tourists did not become significant until January, possibly due partly to the fact that the holiday season travel by local residents were also peaking. According to model sets 1 and 2, *HMP3* also had the highest R^2 for six months (see Table 17), indicating it was the most significant factor among all the variables for these months. Hotel/motel presence also contributed to the decrease in traffic in the summer months of June through October, which were the low season for visitors. Concentration of seasonal residents (*SHP*) also affected traffic volume in a similar fashion as hotels/motels, increasing traffic in February, March, and April while decreasing traffic in the summer months of May through September.

Retired population between 65-75 years of age with the highest income (*RH5_HQ*) seemed to have a strong presence in January and absence in July. This group of population has high disposable income and relative better health compared to older retired population, which allow them to leave the state for vacation during the summer months and return in the winter months.

The retail employment variable, *RETAILP*, was significant in the months of July and August according to model sets 1 and 2 and also in June according to model set 3. It had a negative effect on traffic volume in the summer. It appeared that people reduced their shopping activities in the summer months, which were also the months for vacations, especially July and August.

Table 17 provides the partial R^2's and the significance levels of the variables in the three model sets. The numbers in parentheses indicate the particular months for which the models contained the corresponding variable. Besides *HMP3*, which had the largest R^2s for majority of the months, other variables except *HOTELP* also had reasonable R^2s for certain months, particularly *SHP*, *RH5_HQ*, and *RETAILP*.

Table 15. Best Models Based on Three Buffer Methods for Urban Roads

Month	Model Set 1 (Buffer Method 1 with 5-mile Radius Buffer of Freeway Count Stations)	Adj. R^2	MSE
Jan	$MSF1 = 1.0313 - 0.2959 \times HMP3 - 0.2552 \times RH5\_HQ$	0.7911	0.00051583
Feb	$MSF2 = 0.9913 - 0.2765 \times SHP - 0.3096 \times HMP3$	0.8017	0.00062068
Mar	$MSF3 = 0.98137 - 0.22859 \times SHP - 0.2542 \times HMP3$	0.7681	0.00051201
Apr	$MSF4 = 1.0021 - 0.1581 \times SHP - 0.1060 \times HMP3$	0.6342	0.00028149
May	$MSF5 = 0.9969 + 0.1421 \times SHP$	0.3726	0.00026545
Jun	$MSF6 = 0.9996 + 0.2216 \times SHP + 0.2687 \times HMP3$	0.7748	0.00051171
Jul	$MSF7 = 1.0036 + 0.1659 \times HMP3 + 0.1954 \times RETAILP + 0.1709 \times RH5\_HQ$	0.9033	0.00027782
Aug	$MSF8 = 0.9832 + 0.2093 \times SHP + 0.1838 \times HMP3 + 0.1332 \times RETAILP$	0.8677	0.00035661
Sep	$MSF9 = 1.0227 + 0.1189 \times SHP + 0.2663 \times HMP3$	0.6746	0.00097553
Oct	$MSF10 = 1.0062 + 0.2034 \times HMP3$	0.3060	0.00066651
Nov	$MSF11 = 0.9792 + 1.1637 \times HOTELCAMP$	0.2053	0.00049519
Dec	$MSF12$ = no model	-	-
	Model Set 2 (Buffer Method 2 with 15-mile Service Area for Count Stations on Freeways)	Adj. R^2	MSE
Jan	$MSF1 = 1.0306 - 0.2535 \times HMP3 - 0.2979 \times RH5\_HQ$	0.7637	0.00058340
Feb	$MSF2 = 0.9894 - 0.2775\ SHP - 0.2936 \times HMP3$	0.7551	0.00076649
Mar	$MSF3 = 0.9790 - 0.2046 \times SHP - 0.2664 \times HMP3$	0.7362	0.00058246
Apr	$MSF4 = 1.0013 - 0.1549 \times SHP - 0.1121 \times HMP3$	0.6520	0.00026780
May	$MSF5 = 0.9983 + 0.1369 \times SHP$	0.3400	0.00027924
Jun	$MSF6 = 1.0018 + 0.1964 \times SHP + 0.2847 \times HMP3$	0.7547	0.00055723
Jul	$MSF7 = 1.0042 + 0.1534 \times HMP3 + 0.1879 \times RETAILP + 0.1954 \times RH5\_HQ$	0.8955	0.00030026
Aug	$MSF8 = 0.9833 + 0.2412 \times SHP + 0.1674 \times HMP3 + 0.1256 \times RETAILP$	0.8794	0.00032522
Sep	$MSF9 = 1.0166 + 0.2351 \times SHP + 0.2044 \times HMP3 + 1.5102 \times HOTELP$	0.7467	0.00075935
Oct	$MSF10 = 1.0050 + 0.2151 \times HMP3$	0.3405	0.00081076
Nov	$MSF11$ = no model	-	-
Dec	$MSF12$ = no model	-	-

Table 15. Best Models Based on Three Buffer Methods (Cont.)

	Model Set 3 (Buffer Method 2 with 15-mile Service Area for Count Stations on Freeways)	Adj. R^2	MSE
Jan	$MSF1 = 1.0508 - 0.2830 \times HMP4 - 0.6967 \times RH5\_HQ$	0.6460	0.00087407
Feb	$MSF2 = 0.9987 - 0.3535 \times SHP - 0.3505 \times HMP4$	0.6292	0.00062068
Mar	$MSF3 = 0.9898 - 0.2715 \times SHP - 0.3525 \times HMP4$	0.7233	0.00061090
Apr	$MSF4 = 1.0080 - 0.2081 \times SHP - 0.1590 \times HMP4$	0.7006	0.00023037
May	$MSF5 = 0.9940 + 0.3068 \times RH5\_HQ$	0.3224	0.00028666
Jun	$MSF6 = 0.9414 + 0.2633 \times SHP + 0.3183 \times RETAILP$	0.6272	0.00084718
Jul	$MSF7 = 0.9763 + 0.3523 \times RETAILP + 0.5156 \times RH5\_HQ$	0.7346	0.00076260
Aug	$MSF8 = 0.9554 + 0.3268 \times SHP + 0.3654 \times RETAILP$	0.7518	0.00035661
Sep	$MSF9 = 1.0138 + 0.3572 \times SHP + 0.3185 \times HMP4$	0.5997	0.00120000
Oct	$MSF10 = 1.0004 + 0.2642 \times HMP4$	0.3054	0.00085396
Nov	*MSF11* = no model	-	-
Dec	*MSF12* = no model	-	-

Table 16. Variables and Their Signs for TTMSs on Urban Roads

Variable	Description	Model Set 1	Model Set 2	Model Set 3
HMP3	Ratio of occupied hotel rooms to occupied hotel rooms plus households	Jan, Feb, Mar, Apr (-); Jun, Jul, Aug, Sep, Oct (+)	Jan, Feb, Mar, Apr (-); Jun, Jul, Aug, Sep, Oct (+)	
HMP4	Ratio of hotel population to hotel population plus population			Jan, Feb, Mar, Apr (-); Sep, Oct (+)
SHP	Ratio of seasonal households to permanent households	Feb, Mar, Apr (-); May, Jun, Aug, Sep (+)	Feb, Mar Apr (-); May, June Aug, Sep (+)	Feb, Mar, Apr (-); Jun, Aug, Sep (+);
RH5_HQ	Percentage of retired householders of the highest income quartile	Jan (-); Jul (+)	Jan (-); July (+)	Jan (-); May, Jul, (+)
RETAILP	Retail workers as a percentage of total retail workers plus population	Jul, Aug (+);	July, Aug, (+)	Jun, Jul, Aug (+)
HOTELP	Ratio of hotel employment to the sum of hotel employment and population in buffer area		Sep (+)	
HOTELCAMP	Hotel/RV camp workers as a percentage of hotel/RV camp workers plus population	Nov (+)		

Table 17. Partial R^2's and Significance Levels of Variables from the Three Models

Variable	Description	Model 1		Model 2		Model 3	
		R^2	Pr > |t|	R^2	Pr > |t|	R^2	Pr > |t|
Constant	Model Constant		< 0.0001		< 0.0001		< 0.0001
HMP3	Ratio of occupied hotel rooms to occupied hotel rooms plus households	(1) 0.7196 (2) 0.6942 (3) 0.6669 (4) 0.0873 (6) 0.6834 (7) 0.0695 (8) 0.6898 (9) 0.5807 (10) 0.3327	< 0.0001 < 0.0001 0.0002 0.0200 < 0.0001 0.0019 0.0030 0.0026 0.0016	(1) 0.6689 (2) 0.6570 (3) 0.6663 (4) 0.0915 (6) 0.6928 (7) 0.0671 (8) 0.0442 (9) 0.6308 (10) 0.3659	0.0008 0.0005 0.0003 0.0152 0.0001 0.0061 0.0052 0.0207 0.0008	–	–
HMP4	Ratio of hotel population to hotel population plus population	–	–	–	–	(1) 0.1646 (2) 0.5053 (3) 0.6172 (4) 0.1420 (9) 0.1463 (10) 0.3321	0.0002 0.0021 < 0.0001 0.0018 0.0051 0.0017
SHP	Ratio of seasonal households to permanent households	(2) 0.1228 (3) 0.1190 (4) 0.5750 (5) 0.3967 (6) 0.1087 (8) 0.1333 (9) 0.1189	0.0005 0.0013 0.0023 0.0004 0.0017 0.0009 0.0051	(2) 0.1170 (3) 0.0901 (4) 0.5873 (5) 0.3654 (7) 0.0807 (8) 0.7031 (9) 0.0804	0.0017 0.0065 0.0028 0.0008 0.0074 0.0002 0.0064	(2) 0.1524 (3) 0.1274 (4) 0.5817 (6) 0.1431 (8) 0.1858 (9) 0.4842	0.0033 0.0020 0.0002 0.0042 0.0002 0.0034
RH5_HQ	Percentage of retired householders of the highest income quartile	(1) 0.7196 (7) 0.0876	< 0.001 0.0228	(1) 0.1131 (7) 0.0285	0.0017 0.0139	(1) 0.3485 (5) 0.5087 (6) 0.0935	0.0012 0.0002 0.0058
RETAILP	Retail workers as a percentage of total retail workers plus population	(7) 0.8222 (8) 0.0598	< 0.0001 0.0023	(6) 0.8120 (8) 0.1460	0.0001 0.0027	(6) 0.5127 (7) 0.6615 (8) 0.5851	< 0.0001 < 0.0001 < 0.0001
HOTELP	Ratio of hotel employment to the sum of hotel employment and population in buffer area	–	–	(9) 0.0647	0.0169	–	–
HOTELCAMP	Hotel/RV camp workers as a percentage of hotel/RV camp workers plus population	(11) 0.2358	0.0102	–	–	–	–

No significant factors were identified for the months of November and December, possibly because these were the times when out-of-town visitors and seasonal residents began to stream into Southeast Florida and, at the same time, travel by the locals also increased due to the holiday seasons. Therefore, no single factor was dominant. Lacking other land use variables to predict seasonal factors for these two months, the location dummy variables were introduced into the models. The results are given below, with the variables' partial R^2s and significance presented in Table 18.

Buffer Method 1

$MSF11 = 0.9859 - 0.1001 \times SHP + 1.4776 \times HOTELCAMP + 0.0203 \times DG4 - 0.0460 \times DG5$
(adj. $R^2 = 0.6399$, MSE = 0.00022438)

$MSF12 = 0.9951 - 0.1455 \times SHP - 0.0581 \times DG2 + 0.0306 \times DG4 - 0.0871 \times DG5$
(adj. $R^2 = 0.7226$, MSE = 0.00040620)

Buffer Method 2

$MSF11 = 0.9876 - 0.1044 \times SHP + 1.4131 \times HOTELP + 0.0228 \times DG4 - 0.0626 \times DG5$
(adj. $R^2 = 0.6046$, MSE = 0.00024638)

$MSF12 = 0.9930 - 0.1314 \times SHP + 0.0557 \times DG2 + 0.0283 \times DG4 - 0.0890 \times DG5$
(adj. $R^2 = 0.7003$, MSE = 0.00043874)

Buffer Method 3

$MSF11 = 0.9936 - 0.0486 \times DG5$
(adj. $R^2 = 0.2408$, MSE = 0.00047304)

$MSF12 = 0.9680 + 0.0291 \times DG1 - 0.0355 \times DG2 + 0.0420 \times DG4 - 0.0730 \times DG5$
(adj. $R^2 = 0.7011$, MSE = 0.00043763)

Again, the sign of the variable representing the season population was negative and seemed to indicate that seasonal residents began the annual migration in November. Location dummy variables *DG2* and *DG5* appeared to be stronger predictors than the other location dummy variables. *DG5* indicated locations on the Florida Turnpike. The negative sign of *DG5* indicated an increase in traffic on the turnpike in November and December, possibly due to the increase in holiday related travels. Count stations in the seasonal group represented by *DG2* were not in spatially proximity of each other, and the land use in the surrounding areas also varied significantly. Note that the sign of this variable was inconsistent across models. As a result, it was difficult to interpret the meaning of this dummy variable. Dummy variable *DG4* represents two locations on causeways leading to Miami Beach where tourists and retired population concentrate. The positive sign seemed to indicate that in November and December, traffic on causeways slight decreased. No good explanation could be offered regarding *DG4*.

Table 18. Partial R^2's and Significance Levels for November and December Models

Variable	Description	Buffer Method 1		Buffer Method 2		Buffer Method 3	
		Partial R^2	Pr > |t|	Partial R^2	Pr > |t|	Partial R^2	Pr > |t|
Constant	Model constant		< 0.0001		< 0.0001		< 0.0001
SHP	Ratio of seasonal households to permanent households	(11) 0.1028 (12) 0.1070	0.0007 0.0040	(11) 0.1124 (12) 0.0996	0.0091 0.0101	–	–
HOTELP		(11) 0.2381	< 0.0001	(11) 0.1782	0.0002	–	–
DG1	Dummy variable – located inland	–	–	–	–	(12) 0.0919	0.0098
DG2	Dummy variable – located inland mostly near the western urban boundary	(12) 0.2499	< 0.0001	(12) 0.2499	0.0001	(12) 0.2499	0.0089
DG4	Dummy variable – located on causeways to Barrier Islands (including Miami Beach)	(11) 0.0844 (12) 0.0796	0.0218 0.0122	(11) 0.1048 (12) 0.0682	0.0155 0.0236	(12) 0.0766	0.0026
DG5	Dummy variable – located on Florida Turnpike	(11) 0.2700 (12) 0.3287	0.0004 < 0.0001	(11) 0.2700 (12) 0.3287	< 0.0001 < 0.0001	(11) 0.2700 (12) 0.3287	0.0055 0.0002

5.1.4 Summary

Using the MSFs collected from the TTMS sites in Broward, Miami-Dade, and Broward counties and demographic and socioeconomic data mainly from the census, this research identified several significant factors that contributed to the seasonal patterns of traffic. These factors included the seasonal movement of part-time residents and tourists (through variables that reflect concentration of hotels and motels), retired people between age 65 and 75 with high income, and retail employment. Roadway federal functional classification was not found to be a factor, likely because in large urban areas major roads were used by travel for mixed purposes and no single purpose use was dominant. Similarly, no correlation was found between the seasonal factors and traffic volume per lane and number of lanes. While the results cannot be generalized for other urban areas, they do point to the possibility of determining seasonal factors based on fundamental causes – land uses, demographics, and socioeconomics, which are important determinates of travel demand.

5.2 Rural Area Regression Analysis

Regression analyses were also performed on data from TTMSs in selected Florida rural areas. The objective was to identify possible factors that influenced seasonal factors in Florida rural areas. The differences between the regression analyses for the urban and rural areas mainly lied in the variable definition and buffer methods that were used to determine the variable values. In this section, the study area, variables investigated in the regression analyses, and the regression results are described.

5.2.1 Study Area Selection

Rural roads were defined as those that were not located in an urban area. Table 19 shows the number of TTMSs located on the rural roads in each of the FDOT districts. The GIS coverage files that were available from the 2002 FDOT Traffic Information CD were utilized to classify TTMSs into urban and rural count stations. As shown in Table 19, FDOT Districts 2 and 3 had relatively more rural TTMSs, which allowed more samples to be considered in the subsequent regression analysis. Therefore, these two districts were selected as the study area, where there were 73 TTMSs located on the rural roads with MSFs for every month in a year.

Table 19. Number of Rural Counties and TTMSs

District	1	2	3	4	5	6	7	8	Total
Number of TTMSs	23	30	43	9	16	1	5	2	129

5.2.2 Study Data

The independent variables prepared to calibrate the multiple regression models for the rural TTMSs included roadway characteristics, demographic and socioeconomic variables, and other variables that described the location and accessibility of the TTMSs. The following sections describe the independent variables that were compiled for each TTMS on the rural roads in FDOT Districts 2 and 3.

5.2.2.1 *Roadway Characteristic Variables*

Variables in this category are given in Table 20, where Variables *Fc1*, *Fc2*, *Fc6*, *Fc7,* and *Fc8* were dummy variables indicating the type of road where a TTMS was located. Variables *DirNS* and *DirEW* were dummy variables specifying the orientations of a given roadway's alignment. The data were retrieved from the 2002 FDOT Traffic Information CD and the FDOT's Roadway Characteristics Inventory (RCI) database.

Table 20. Roadway Characteristic Variables for Rural Roads

Variable	Description
Lanes	Number of lanes
Tfctr	Truck factor
Fc1	1if TTMS was located on freeway; 0 otherwise
Fc2	1 if TTMS was located on principal arterial; 0 otherwise
Fc6	1 if TTMS was located on minor arterial; 0 otherwise
Fc7	1 if TTMS was located on major collector; 0 otherwise
Fc8	1 if TTMS was located on minor collector; 0 otherwise
DirNS	1 if roadway runs in the north-south direction; 0 otherwise
DirEW	1 if roadway runs in the east-west direction; 0 otherwise

5.2.2.2 *Demographic and Socioeconomic Variables*

Most of the demographic and socioeconomic data for rural area were retrieved from the 2000 Census with the exception of the employment data, which were from a proprietary database purchased by FDOT. This assured the data availability to all FDOT districts. Two buffer methods were used to compile the data for the demographic and socioeconomic variables. The first buffer method employed a set of buffer sizes that varied according to the functional classifications of the roadway segments on which the TTMSs were located. A circular buffer around each count station was created to define the impact area and the associated demographic and socioeconomic data were then aggregated. Two sets of buffer radii were used as shown in Table 21.

Table 21. Buffer Sizes Based on Functional Classification

Data Set	Variable	Description	Buffer Size (miles)
1	*Fc1*	Rural Principal Arterial — Interstate	15
	Fc2	Rural Principal Arterial — Other	12
	Fc6	Rural Minor Arterial	8
	Fc7	Rural Major Collector	3
	Fc8	Rural Minor Collector	1
2	*Fc1*	Rural Principal Arterial — Interstate	10
	Fc2	Rural Principal Arterial — Other	8
	Fc6	Rural Minor Arterial	5
	Fc7	Rural Major Collector	2
	Fc8	Rural Minor Collector	1

As mentioned in the previous section, a larger buffer zone implied that the MSFs for a count station were impacted by the characteristics of a larger surrounding area. For the TTMSs on the rural roads larger buffer sizes than those used in the development of the urban models were applied to incorporate land use effects from distant developments.

The second buffer method defined the impact area based on the average travel time to workplace from Census 2000. Since trips to and from work are one of the major activities that determine the traffic generated or attracted by an area, the average travel time to workplace may suggest how far people need or are willing to travel to work. Based on Census 2000, the average travel time to workplace for the counties in FDOT Districts 2 and 3 was 28 minutes. To reduce the computation complexity, a 30-minute average travel time was assumed and the corresponding travel distances from a given TTMS following the road network were calculated based on the posted speed limits to estimate its impact area. Using the above two buffer methods, the following independent variables were compiled and three datasets were created:

Rural Population

Variable *Pop* was the population in the buffer zone around a TTMS.

Population Density

Variable *Popden* was the population density in the buffer zone around a TTMS.

Population by Age Groups

The population in a buffer zone of a TTMS was divided into six age groups to capture the potential effects it might have on seasonal traffic patterns. The variables are given in Table 22.

Table 22. Population Age Group Variables

Variables	Description
Ageunder5	Population density aged 5 and under
Age5_17	Population density aged between 5 and 17
Age22_65	Population density aged between 22 and 65
Age18_64	Population density aged between 18 and 64
Age5_21	Population density aged between 5 and 21
Age18_21	Population density aged between 18 and 21
Age65up	Population density aged 65 and over
Punder5	Population aged 5 and under as a percentage of total population
P5_17	Population aged between 5 and 17 as a percentage of total population
P22_65	Population aged between 22 and 65 as a percentage of total population
P18_64	Population aged between 18 and 64 as a percentage of total population
P5_21	Population aged between 5 and 21 as a percentage of total population
P18_21	Population aged between 18 and 21 as a percentage of total population
P65up	Population aged 65 and over as a percentage of total population

Households and Seasonal Household Percentage

The *Ruhh* variable represented the number of households within a buffer area. *Sshh* was the total while *Sshden* the density of seasonal households located in a buffer area. Variable *P_Shhphh* was the seasonal households as a percentage of permanent households and *P_Shhthh* the seasonal households as a percentage of total households in the buffer zone around a count station.

Hotel/Motel Rooms

Hrooms was the number of hotel/motel rooms in the buffer zone around a count station.

Employment Variables

Table 23 defines the 13 employment variables developed in this study.

Table 23. Employment Variable Definitions for Rural Roads

Variable	Description
T_Employ	Number of total employment within a buffer zone
T_Empden	Employment density around a count station
He_Te	Hotel employment as a percentage of total employment
C_Employ	Employment for crop production as a percentage of total employment
Ag_Emplo	Employment for livestock production as a percentage of total employment
Craged	Employment for agricultural (crop and livestock) production as a percentage of total employment
E_Ce	Employment density not including crop employment
E_Craged	Employment density not including crop and livestock employment
Co_Te	Commercial employment as a percentage to total employment
In_Te	Industrial employment as a percentage to total employment
Se_Te	Service employment as a percentage to total employment
Pop_Pe	$Pop\_Pe = \text{population in a buffer} / (\text{total employment} \times \text{population})$
Emp_Pe	$Emp\_Pe = \text{employment in a buffer} / (\text{total employment} \times \text{population})$

Income Variables

Variable *A_Income* was the arithmetic average of the average income at the census block group level in the buffer zone around a TTMS. *Aincahh* was the average income weighted by the number of households of the area, as given by the following equation:

$$Aincahh = \sum \text{Average income} \times \text{HHs} \ / \text{Number of total HHs} \quad (50)$$

5.2.2.3 *Other Variables*

The following variables described the network density, accessibility to areas of different interest from the TTMSs, and TTMS locations.

TTMS Relative Locations

Fifteen variables described the distance from a TTMS to coastlines, metropolitan areas, state borders, and highway interchanges. Table 24 defines the variables in this category.

Table 24. Position Variables

Variable	Description
Distcoast	Shortest distance between a TTMS and the coast line
Sqdcoast	Square root of DISTCOAST
Ra_Distsq	$1/Distcoast^2$
Dist1	Maximum of $\frac{\text{population of a given metropolitan area}}{\text{distance between a TTMS and the metropolitan area}}$
Distmetr	Shortest distance between a TTMS and the closest metropolitan area
Sqrdistmet	Square root of *Distmetr*
Ra_Distmet	$1/Distmetr^2$
Distbor	Shortest distance from a TTMS to the state line between Florida and Georgia or Alabama
Sqdbor	Square root of *Distbor*
Ra_Disborsq	$1/Distbor^2$
Interdist	Distance from a TTMS to the closest highway interchange
Sqrinterdi	Square root of *Interdist*
Ra_Interdistsq	$1/Interdist^2$
Indexd1	$\sum \frac{\text{Population for a metropolitan area}}{(\text{Distance from the TTMS to a metropolitan area})^2}$
Indixdist2	$1/Indexd1$

Ramp Access

Ramp was a dummy variable indicating whether a TTMS lied on a roadway segment with direct access to Interstate I-10. A value of 1 indicated a direct connection to a ramp and 0 otherwise.

Roadway Coverage

Variable *Rlength* was the sum of the lengths of all roadway segments within the buffer of a TTMS. Variable *Rdensity* represented roadway density and was calculated for each buffer as follows:

$$Rdensity = \sum(\text{Length of the roadway segment} \times \text{Number of lanes}) / \text{Buffer area} \qquad (51)$$

Geographic Locations

Six dummy variables, *G2*, *G3*, *G5*, *G6*, *G7*, and *G8*, were defined based on the results from the model-based cluster analysis that considered both seasonal traffic fluctuations and locations of TTMSs. The results from the EII model with six seasonal factor groups were selected to define these dummy variables. The model and the number of groups were selected because of the relatively larger area covered by the six factor groups, as indicated by the circles drawn around the group centroids (see Figure 17).

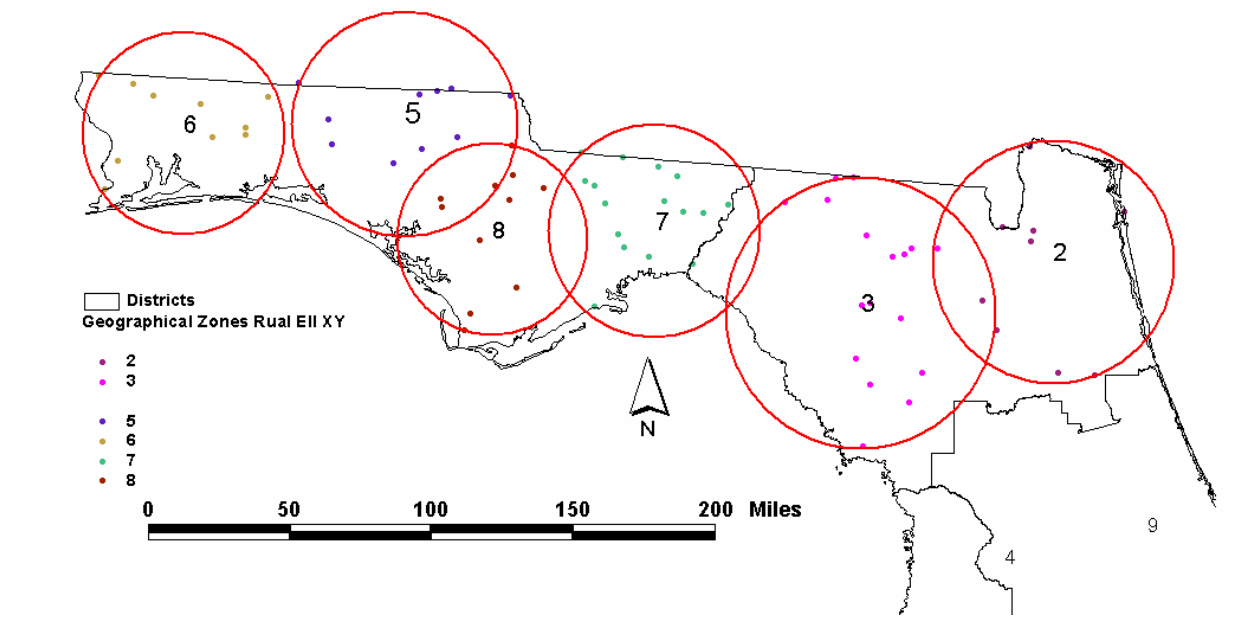


Figure 17. The Spatial Extent of the Six Factor Groups Used to Compile Location Parameters

5.2.3 Multiple Linear Regression Analysis

Three datasets were compiled, each based on a specific set of buffer sizes as described in Section 5.2.2.2. Stepwise regression was applied in the variable selection procedure. Significant levels were set at 0.05 for regressors to enter and stay in a model. Multicollinearity among independent variables was checked based on VIFs, which should be less than 10. The dummy variables for geographical locations were first excluded to test the explanatory power of the land use variables. The results based on the three data sets after excluding the geographical location variables are shown in Tables 25 through 27, respectively, where the models with the highest R^2's among the three model sets are highlighted. The results showed that other explanatory variables needed to be considered in the calibration process because much lower R^2's were obtained from the TTMSs located on rural roads than those on urban roads, except for May, September, and December.

The same regression analysis was performed again by including the geographic location dummy variables *G2*, *G3*, *G5*, *G6*, *G7*, and *G8* in the calibration process. The results are shown in Tables 28 through 30, where the models with the highest R^2's among the three model sets calibrated using the three data sets are highlighted. Tables 28 to 30 show that improved R^2's were obtained from the models for April, July, and September when compared with the models in Tables 25 to 27. However, the improvements did not appear to be significant. Table 31 summarizes the variables of the three sets of models with the signs of their coefficients indicated in the parentheses. It may be seen that the variables and their signs were mostly consistent across the three sets of models.

Table 32 provides the partial R^2 and significance level for each of the variables entered into the models. A higher R^2 is indicative of the explanatory power of a variable. Variables that appeared in multiple months and multiple models such as *Age65up*, *E_Ce*, *Fc1*, *He_Te*, *P_Shhthh*, and *Tfctr* were likely to be important to explaining seasonal traffic patterns. For instance, traffic on I-10 increased in March and July, as suggested by at least two models, which could be associated with spring break and summer vacation traffic. The truck factor variable, *Tfctr*, indicated that in November and December traffic volume went up, which was likely a result of increased freight activities for the holiday season. Variable *Age65up*, which represented retired population, seemed to indicate that traffic in the colder months picked up but dropped in hotter months, possibly because many of them spent time in the winter in Florida but moved back to the north in the spring. The hotel employment variable, *He_Te*, might have suggested that traffic increased in March, possibly because of spring breakers. *P_Shhthh*, which represented seasonal households, seemed to suggest that traffic increased in the spring and summer but decreased in the fall and winter. This is the opposite of what happens in Southeast Florida. However, the explanation might lie in the difference between the seasonal populations in south and north Florida. While the seasonal residents in south Florida are more likely to be retired people and "snow birds", the seasonal residents in north Florida might be mostly agricultural workers.

To better understand what affect the seasonal change in traffic as well as the variables that have been identified, more information and detailed analysis is needed regarding the economic activities in north Florida.

Table 25. Models for Buffer Size 1 without Geographical Location Variables

Month	Model Set 1 (Buffer size = 15, 12, 8, 3, 1)	Adj. R^2	MSE
Jan	*MSF1* = 1.15311 – 0.00483 × *Age65up* – 0.46272 × *Ag_Emplo*	0.0941	0.00427
Feb	*MSF2* = 1.07552 – 0.00564 × *Age65up* – 0.09376 × *Hrooms*	0.1605	0.00315
Mar	*MSF3* = 0.94826 – 1.04928 × *He_Te* – 0.04729 × *Fc1* + 0.01058 × *Sqdcoast*	0.3153	0.00141
Apr	*MSF4* = 0.97266 + 0.00245 × *Age65up* – 0.00193 × *P_Shhthh* + 0.04064 ×Fc1 – 55.84843 × *Dist1*	0.4191	0.00108
May	*MSF5* = 0.94323 – 0.00283 × *P_Shhthh* + 0.85691 × *He_Te* + 0.001 × *E_Ce* + 0.00188 × *Tfctr*	0.5402	0.00082
Jun	*MSF6* = 0.85688 + 0.00216 × *E_Ce* + 0.0000021 × *A_Income* + 0.07653 × *Hrooms*	0.2755	0.00241
Jul	*MSF7* = 0.94707 + 0.00253 × *E_Ce* – 0.0875 × *Fc1* – 0.00428 ×*P_Shhthh* + 1.26438 × *He_Te*	0.4066	0.00420
Aug	*MSF8* = 1.027 – 0.58599 × *Punder5* + 0.82631 × *He_Te* – 0.00124 × *P__Shhthh*	0.1568	0.00133
Sep	*MSF9* = 1.03839 + 0.12371 × *Fc1* + 0.00327 × *P__Shhthh* – 0.0011 × *E_Ce*	0.5255	0.00245
Oct	*MSF10* = 0.99876 – 0.00125 × *E_Craged* – 0.00023887 × *Ra_Disborsq* + 0.00247 × *P_Shhthh* + 0.05087 × *Fc1* + 0.47028 × *Ag_Emplo*	0.4909	0.00164
Nov	*MSF11* = 1.07546 + 0.00184 × *E_Craged* – 0.00241 × *Tfctr*	0.2438	0.00267
Dec	*MSF12* = 1.09384 – 0.00012146 × *Rdlength* + 0.00334 × *P_Shhthh* – 0.00392 × *Tfctr* – 0.00151 × *E_Ce*	0.4935	0.00370

Table 26. Models for Buffer Size 2 without Geographical Location Variables

Month	Model Set 2 (Buffer size = 10, 8, 5, 2, 1)	Adj. R^2	MSE
Jan	*MSF1* = 1.13830 + 0.00299 × *P_Shhthh* – 0.01019 × *Sqdbor*	0.1005	0.00414
Feb	*MSF2* = 1.05738 – 0.00412 × *Age65up*	0.10811	0.00345
Mar	*MSF3* = 0.96247 – 0.00021148 × *Rdlength* – 0.00501 × *Sqdbor* + 0.00210 × *Age65up* + 0.1179 × *Sqdcoast* – 0.45976 × *He_Te*	0.3747	0.00129
Apr	*MSF4* = 0.98425 – 0.00193 × *P_Shhthh* + 0.03965 ×Fc1 – 46.43197 × *Dist1*	0.4191	0.00108
May	*MSF5* = 0.94983 – 0.00235 × *P_Shhthh* + 0.00209 × *Tfctr* + 0.00057520 × *E_Ce*	0.5493	0.00084
Jun	*MSF6* = 0.88358 + 0.00467 × *Age65up* + 0.00000160 × *A_Income* – 0.03557 × *Fc1*	0.2213	0.00260
Jul	*MSF7* = 0.97191 + 0.00133 × *E_Ce* – 0.09289 × *Fc1* – 0.00585 × *P_Shhthh*	0.3604	0.00420
Aug	*MSF8* = 0.96903 – 0.00221 × *Tfctr*	0.1692	0.00131
Sep	*MSF9* = 1.02284 + 0.12868 × *Fc1* + 0.0504 × *P__Shhthh*	0.4771	0.00270
Oct	*MSF10* = 1.01150 + 0.00212 × *P_Shhthh* + 0.04887 × *Fc1* – *0.00029792* × Ra_Disborsq – 0.00074420 × *E_Craged*	0.4217	0.00186
Nov	*MSF11* = 1.05344 + 0.00188 × *P_Shhthh* – 0.00098311 × *E_Craged* – 0.00221 × *Tfctr*	0.2481	0.00266
Dec	*MSF12* = 1.10180 + 0.00368 × *P_Shhthh* – 0.00638 × *Tfctr* – 0.00564 × Age65up	0.5581	0.00322

Table 27. Models for Buffer Size 3 without Geographical Location Variables

Month	Model Set 3 (Buffer size = 30-minute driving time)	Adj. R^2	MSE
Jan	*MSF1* = 0.99803 + 55.24827 × *Indixdist2* – 0.00379 × *Age5_17*	0.1358	0.00410
Feb	*MSF2* = 1.08452 – 0.03208 × *Dir* – 0.00452 × *E_Ce* – 0.01380 × *Sqdbor* + 0.00452 × *P_shhthh*	0.2562	0.00280
Mar	*MSF3* = 0.96420 – 0.04575 × *Fc1* – 0.00139 × *Distborder* + 0.64524 × *Craged* + 0.00285 *P_shhthh*	0.3832	0.00129
Apr	*MSF4* = 0.95770 – 0.03698 × *Fc1* + 0.52329 × *Craged* – 44.31749 × *Dist1* – 0.00123 × *Distbor* + 0.0000353 × *Sshh*	0.4724	0.00099
May	*MSF5* = 0.95615 + 0.00000190 × *T_Employ* – 0.00359 × *P_Shhthh*	0.3216	0.00127
Jun	*MSF6* = 0.92787 + 0.06376 × *Fc1* + 0.00000324 × *T_employ*	0.1921	0.00273
Jul	*MSF7* = 0.93106 + 0.00000378 × *T_Employ* – 0.11981 × *Fc1* – 0.00354 × *P_Shhthh*	0.2873	0.00510
Aug	*MSF8* = 0.96874 – 0.00222 × *Tfctr*	0.1792	0.00133
Sep	*MSF9* = 1.01432 + 0.4287 × *Fc1* + 0.00218 × *Interdist*	0.4444	0.00291
Oct	*MSF10* = 1.06765 – 1.03873 × *Craged* + 0.06055 × *Fc1* – 0.00000227 × *T_Employ*	0.2974	0.00228
Nov	*MSF11* = 1.06032 + 0.00000255 × *T_Employ*	0.1086	0.00319
Dec	*MSF12* = 1.13767 + 0.00359 × *P_Shhthh* – 0.00008806 × *Tfctr*	0.3695	0.00465

Table 28. Models for Buffer Size 1 with Geographical Location Variables

Month	Model Set 1 (Buffer size = 15, 12, 8, 3, 1)	Adj. R^2	MSE
Jan	*MSF1* = 1.15311 – 0.00483 × *Age65up* – 0.46272 × *Ag_Emplo*	0.0941	0.00427
Feb	*MSF2* = 1.07552 – 0.00564 × *Age65up* – 0.09376 × *Hrooms*	0.1605	0.00315
Mar	*MSF3* = 0.94826 – 1.04928 × *He_Te* – 0.04729 × *Fc1* + 0.01058 × *Sqdcoast*	0.3153	0.00141
Apr	*MSF4* = 0.97266 + 0.00245 × *Age65up* – 0.00193 × *P_Shhthh* + 0.04064 ×Fc1 – 55.84843 × *Dist1*	0.4191	0.00108
May	*MSF5* = 0.94323 – 0.00283 × *P_Shhthh* + 0.85691 × *He_Te* + 0.001 × *E_Ce* + 0.00188 × *Tfctr*	0.5402	0.00082
Jun	*MSF6* = 0.85688 + 0.00216 × *E_Ce* + 0.0000021 × *A_Income* + 0.07653 × *Hrooms*	0.2755	0.00241
Jul	*MSF7* = 0.94707 + 0.00253 × *E_Ce* – 0.0875 × *Fc1* – 0.00428 ×*P_Shhthh* + 1.26438 × *He_Te*	0.4066	0.00420
Aug	*MSF8* = 1.027 – 0.58599 × *Punder5* + 0.82631 × *He_Te* – 0.00124 × *P__Shhthh*	0.1568	0.00133
Sep	*MSF9* = 1.02050 + 0.13117 × *Fc1* + 0.00342 × *P__Shhthh* – 0.06591 × *G2*	0.5488	0.00233
Oct	*MSF10* = 0.99876 + 0.47154 × *E_Craged* – 0.00023885 × *Ra_Disborsq* + 0.00247 × *P_Shhthh* + 0.05087 × *Fc1* + 0.47029 × *E_Ce*	0.4909	0.00164
Nov	*MSF11* = 1.07546 + 0.00184 × *E_Craged* – 0.00241 × *Tfctr*	0.2438	0.00267
Dec	*MSF12* = 1.09384 – 0.00012146 × *Rdlength* + 0.00334 × *P_Shhthh* – 0.00392 × *Tfctr* – 0.00151 × *E_Ce*	0.4935	0.00370

Table 29. Models for Buffer Size 2 with Geographical Location Variables

Month	Model Set 2 (Buffer size = 10, 8, 5, 2, 1)	Adj. R^2	MSE
Jan	*MSF1* = 1.13830 + 0.00299 × P_Shhthh – 0.01019 × *Sqdbor*	0.1005	0.00414
Feb	*MSF2* = 1.05738 – 0.00412 × *Age65up*	0.10811	0.00345
Mar	*MSF3* = 0.96247 – 0.00021148 × *Rdlength* – 0.00501 × *Sqdbor* + 0.00210 × *Age65up* + 0.1179 × *Sqdcoast* – 0.45976 × *He_Te*	0.3747	0.00129
Apr	*MSF4* = 0.98425 – 0.00193 × *P_Shhthh* + 0.03965 × *Fc1* – 46.43197 × *Dist1*	0.4191	0.00108
May	*MSF5* = 0.94983 – 0.00235 × *P_Shhthh* + 0.00209 × *Tfctr* + 0.00057520 × *E_Ce*	0.5493	0.00084
Jun	*MSF6* = 0.88358 + 0.00467 × *Age65up* + 0.00000160 × *A_Income* – 0.03557 × *Fc1*	0.2213	0.00260
Jul	*MSF7* = 0.97191 + 0.00133 × *E_Ce* – 0.09289 × *Fc1* – 0.00585 ×*P_Shhthh*	0.3604	0.00420
Aug	*MSF8* = 0.96903 – 0.00221 × *Tfctr*	0.1692	0.00131
Sep	*MSF9* = 1.01894 + 0.13317 × *Fc1* + 0.00476 × *P__Shhthh* + 0.06712 × *G2*	0.5271	0.00244
Oct	*MSF10* = 0.99524 + 0.00255 × *P_Shhthh* + 0.05127 × *Fc*1 – 0.00074420 × *E_Craged* + 0.20532 × *E_Ce*	0.3729	0.00202
Nov	*MSF11* = 1.05344 + 0.00188 × *P_Shhthh* – 0.00098311 × *E_Craged* – 0.00221 × *Tfctr*	0.2481	0.00266
Dec	*MSF12* = 1.10180 + 0.00368 × *P_Shhthh* – 0.00638 × *Tfctr* – 0.00564 × *Age65up*	0.5581	0.00322

Table 30. Models for Buffer Size 3 with Geographical Location Variables

	Model Set 3 (Buffer size = 30-minute driving time)	Adj. R^2	MSE
Jan	$MSF1 = 0.99803 + 55.24827 \times Indixdist2 - 0.00379 \times Age5\_17$	0.1358	0.00410
Feb	$MSF2 = 1.08452 - 0.03208 \times Dir - 0.00452 \times Et\_Ce - 0.01380 \times Sqdbor + 0.00452 \times P\_shhthh$	0.2562	0.00280
Mar	$MSF3 = 0.96420 - 0.04575 \times Fc1 - 0.00139 \times Distbor + 0.64524 \times Craged + 0.00285 \times P\_shhthh$	0.3832	0.00129
Apr	$MSF4 = 0.95770 - 0.03698 \times Fc1 + 0.52329 \times Craged - 44.31749 \times Dist1 - 0.00123 \times Distbor + 0.0000353 \times Sshh$	0.4724	0.00099
May	$MSF5 = 0.95615 + 0.00000190 \times T\_Employ - 0.00359 \times P\_Shhthh$	0.3216	0.00127
Jun	$MSF6 = 0.92787 + 0.06376 \times Fc1 + 0.00000324 \times T\_employ$	0.1921	0.00273
Jul	$MSF7 = 0.93106 + 0.00000378 \times T\_Employ - 0.11981 \times Fc1 - 0.00354 \times P\_Shhthh$	0.2873	0.00510
Aug	$MSF8 = 0.96587 + 0.00223 \times Tfctr + 0.03931 \times G2$	0.2231	0.00125
Sep	$MSF9 = 1.01125 + 0.14626 \times Fc1 + 0.00202 \times Interdist + 0.00202 \times G2$	0.4942	0.00265
Oct	$MSF10 = 1.06982 - 0.97332 \times Craged + 0.07556 \times Fc1 - 0.00000317 \times T\_Employ + 0.07422 \times G2$	0.3887	0.00199
Nov	$MSF11 = 1.06032 + 0.00000255 \times T\_Employ$	0.1086	0.00319
Dec	$MSF12 = 1.13767 + 0.00359 \times P\_Shhthh - 0.00008806 \times Tfctr$	0.3695	0.00465

Table 31. Variables from the Three Models and the Signs of Their Coefficients

Variable	Description	Model 1	Model 2	Model 3
Employment Variables				
Craged	Employment for agricultural (crop and livestock) production as a percentage of total employment			(+) Mar, Apr (–) Oct
Ag_Emplo	Employment for livestock production as a percentage of total employment	(–) Jan (+) Oct		
E_Ce	Employment density not including crop employment	(–) Sep, Dec (+) May, Jun, Jul	(+) May, Jul,	(–) Feb
E_Craged	Employment density not including crop and livestock employment	(–) Oct (+) Nov	(–) Oct, Nov	
He_Te	Hotel employment as a percentage of total employment	(+) May, Jul, Aug (–) Mar	(–) Mar	
T_Employ	Number of total employment within a buffer zone			(–) Oct (+) May, Jun, Jul, Nov
Population Variables				
Age65up	Population density within the ages 65 and up	(+) Apr (–) Jan, Feb	(+) Jun (–) Feb, Mar, Dec	
P5_17	Population density within the ages 5 to 17			(–) Jan
A_Income	Arithmetic average of the average income	(+) Jun	(+) Jun	
Punder5	Population within the ages 5 and under as a percentage of total population	(–) Aug		
Roadway Characteristics				
DirNS	Roadway runs in the north – south direction			(–) Feb
Fc1	1if TTMS located on freeway; 0 otherwise	(+) Apr, Sep, Oct (–) Mar, Jul	(+) Apr, Sep, Oct (–) Jun, Jul	(+) Jun, Sep, Oct (–) Mar, Jul, Apr
Rdlength	The sum of the length of all roadway segments contained in the buffer of each TTMS	(-) Dec	(–) Mar	

Variable	Description	Model 1	Model 2	Model 3
Tfctr	Truck factor	(–) Nov, Dec (+) May	(–) Aug, Nov, Dec (+) May	(–) Aug, Dec
Seasonal Household Variables				
Hrooms	The number of hotel/motel rooms around a count station	(+) Jun (–) Feb		
P_Shhthh	The seasonal households as a percentage of total households	(+) Sep, Oct, Dec (–) Apr, May, Jul, Aug	(+) Jan, Sep, Oct, Nov, Dec (–) Apr, May, Jul	(+) Feb, Mar, Dec (–) May, Jul
Sshh	The number of seasonal households			(+) Apr
Location Variables				
Dist1	Max of ratio of Population of a metropolitan area to the distance from the TTMS to the Metropolitan area	(–) Apr	(–) Apr	(–) Apr
Distbor	Shortest distance from a TTMS to the state line between Florida and Georgia or Alabama			(–) Mar, Apr
Indexdist2	$\left(\sum \frac{\text{Metropolitan Population}}{\text{Distance from the TTMS to the metropolitan area}}\right)^{-1}$			(–) Jan
Interdist	Distance from a TTMS to the closest highway interchange			(+) Sep
Sqdbor	Square root of shortest distance from a TTMS to the state line between Florida and Georgia or Alabama		(–) Jan, Mar	(–) Feb
Sqdcoast	Square root of shortest distance between a TTMS and the coast line	(+) Mar	(+) Mar	
Radistbodsq	$1/Distbor^2$	(–) Oct	(–) Oct	

Table 32. Partial R2 and Significance Level of Parameters in the Three Rural Model

Variable	Model 1			Model 2			Model 3		
	Month	Par. R^2	Pr. > F	Month	Par. R^2	Pr. > F	Month	Par. R^2	Pr. > F
A_Income	6	0.0674	0.0151	6	0.0516	0.0323			
Ag_Emplo	1	0.0517	0.0465						
	10	0.0722	0.0021						
Age5_17							1	0.0542	0.0365
Age65up	1	0.0675	0.0264	2	0.0939	0.0084			
	2	0.0942	0.0083	3	0.0669	0.0120			
	4	0.0389	0.0316	6	0.1454	0.0009			
				12	0.0798	0.0006			
Craged							3	0.0733	0.0067
							4	0.1032	0.0024
							10	0.1722	0.0003
Dir							2	0.0623	0.0332
Dist1	4	0.1377	0.0001	4	0.1264	0.0007	4	0.0613	0.0141
Distbor							3	0.1428	0.0004
							4	0.0747	0.0046
E_Ce	5	0.0719	0.0013	5	0.0593	0.0036	2	0.0642	0.0264
	6	0.1731	0.0003						
	7	0.1830	0.0002	7	0.1319	0.0016			
	9	0.0307	0.0344						
	12	0.0398	0.0202						
E_Tcraged	10	0.1672	0.0003	10	0.0915	0.0012			
	11	0.1063	0.0049	11	0.0555	0.0303			
FC1	3	0.1381	0.0012				3	0.1381	0.0012
	4	0.169	0.0003	4	0.1132	0.0005	4	0.169	0.0003
				6	0.0568	0.0288	6	0.1386	0.0008
	7	0.0996	0.0026	7	0.1011	0.0034	7	0.1105	0.0041
	9	0.3314	<.0001	9	0.3314	<.0001	9	0.3314	<.0001
	10	0.0925	0.0011	10	0.1406	0.0004	10	0.0645	0.0176
He_Te	3	0.1482	0.0003	3	0.064	0.0084			
	5	0.079	0.0016						
	7	0.0494	0.0169						
	8	0.0536	0.0410						
Hrooms	2	0.0896	0.0071						
	6	0.0652	0.0131						
Indixdist2							1	0.1107	0.0040
Interdist							9	0.1277	0.0001
P_Shhthh				1	0.0548	0.0462	2	0.1055	0.0020
							3	0.0531	0.0162
	4	0.1058	0.0021	4	0.1723	0.0003			
	5	0.3085	<.0001	5	0.3913	<.0001	5	0.1034	0.0018
	7	0.1075	0.0009	7	0.1522	<.0001	7	0.0547	0.0215
	8	0.0581	0.0291	9	0.1545	<.0001			
	9	0.1832	<.0001	10	0.1404	0.0011			
	10	0.0598	0.0132	11	0.1496	0.0007			
	12	0.1402	0.0001	12	0.3235	<.0001	12	0.0729	0.0079
Punder5	8	0.0802	0.0152						

Radisborsq	10	0.1345	0.0005						
Rdlength	12	0.2708	<.0001						
Shh							4	0.0822	0.0016
Sqdbord				1	0.0901	0.0083	2	0.0732	0.0143
				3	0.0905	0.0052			
Sqdcoast	3	0.0575	0.0164	3	0.0478	0.0282			
T_Employ							5	0.2179	<.0001
							6	0.0776	0.0170
							7	0.1540	0.0003
							10	0.0805	0.0057
							11	0.1193	0.0028
Tfctr	5	0.1064	0.0007	5	0.0984	0.0005			
				8	0.1808	0.0002	8	0.1808	0.0002
	11	0.0945	0.0053	11	0.0743	0.0095			
	12	0.0709	0.0030	12	0.1732	<.0001	12	0.2461	<.0001

5.2.4 Summary

It appears that the buffer methods used on rural TTMSs were not adequate to capture the effects of land use and the spatial structure of activities that cause significant traffic fluctuations over time on rural roads. For example, the effect of through traffic, which was not originated or destined in local buffer areas, might not have been well reflected in the variables compiled for the area surrounding a given TTMS. Because the higher the function class of a road is, the more significant through traffic will be, especially on rural roads, the through traffic may be a significant factor. New variables may also need to be developed to better quantify the impact of land use as well as socioeconomic/demographic factors on roadway traffic. However, the regression models did indicate that variables such as functional classification and percentage of seasonal households were significant with relatively high impacts on seasonal traffic patterns. These variables could be used to assist in assigning factor groups to PTMSs.

5.3 Grouping and Assignment Procedures

This section first presents a procedure to appropriately group TTMSs into SF categories. The groupings were originally created via model-based cluster analysis and then fine-tuned to better reflect the spatial fluctuations in the estimation of traffic volumes. Based on the grouping results and the four land use attributes (as discussed in Sections 4.2 and 5.1) observed at the TTMSs for each SF group, a fuzzy decision tree was constructed and applied to determine the SF category for a given PTMS. Application of the grouping and assignment procedures to the tri-county area of southeast Florida was also described.

5.3.1 Grouping Procedure

As described in Section 4.2, the results from the parametric model-based clustering analysis (e.g., the EII model) may be used as a starting point to determine seasonal factor categories by simultaneously considering the spatial proximity and seasonal traffic fluctuations. The grouping procedure is summarized as follows:

Step 1: Perform model-based clustering analysis to determine initial seasonal factor groups.

Step 2: Examine individual seasonal factor groups to identify any TTMSs that do not belong to their original groups because of a different seasonal pattern.

Step 3: Reassign or create new groups as necessary for the TTMSs identified in Step 2 based on their seasonal profiles.

For illustration purposes, consider the Tri-County area in the southeast Florida region, which include Broward, Miami-Dade, and Palm Beach counties. The model-based clustering analysis, which considered the locations of the TTMSs, produced four seasonal factor groups in the tri-county area based on the 2002 statewide monthly seasonal factor data. As shown in Figure 18, these groups were Factor Groups 1, 3, 7, and 21, which were renumbered as 1, 2, 3, and 4, respectively, in the discussion that follows. These new numbers are plotted in red outside the circles that indicate the spatial extent of each group in Figure 18. Figures 19 to 22 show the MSFs for the TTMSs in the same group along with the mean MSFs and thresholds defined as ±10% of the mean MSFs for the four factor groups.

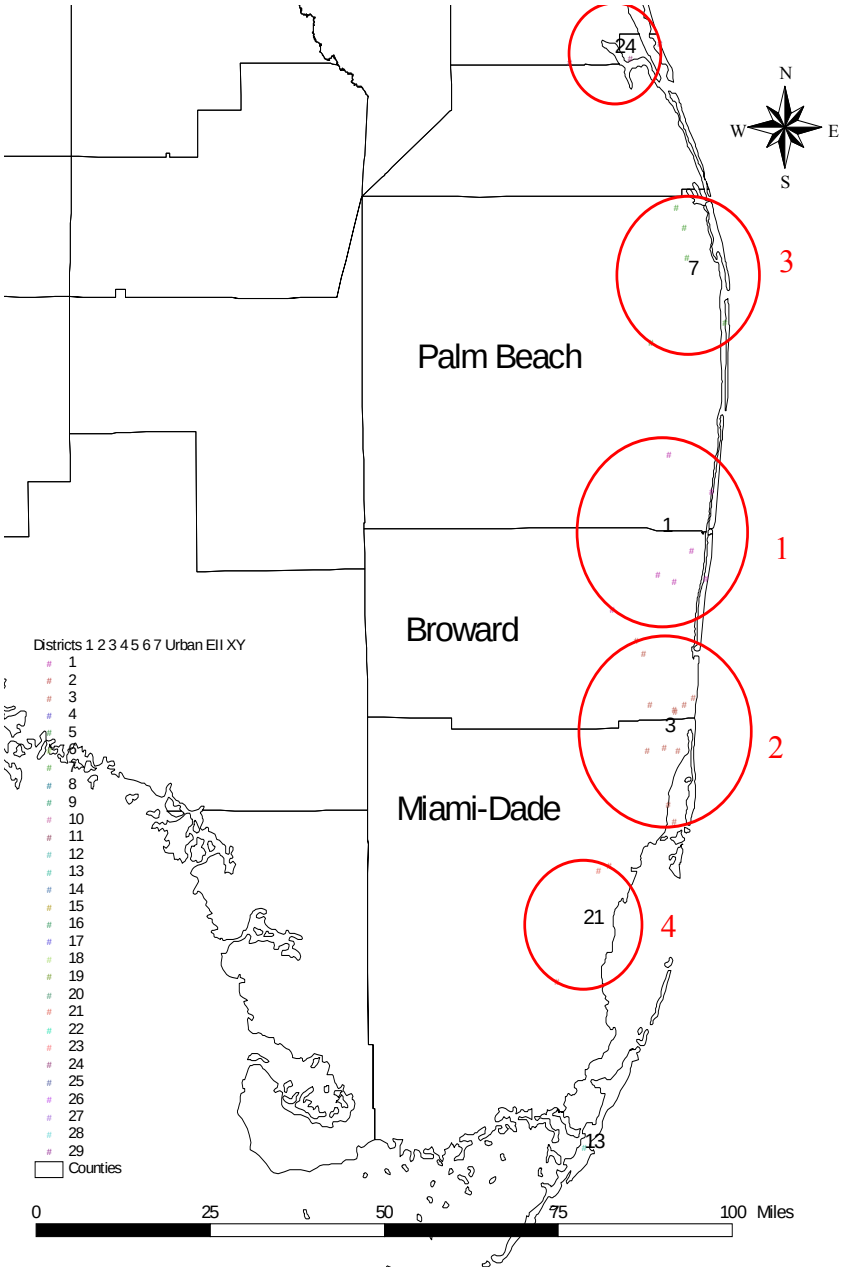


Figure 18. SF Categories from EII Model by Simultaneously Considering TTMS Locations with MSFs

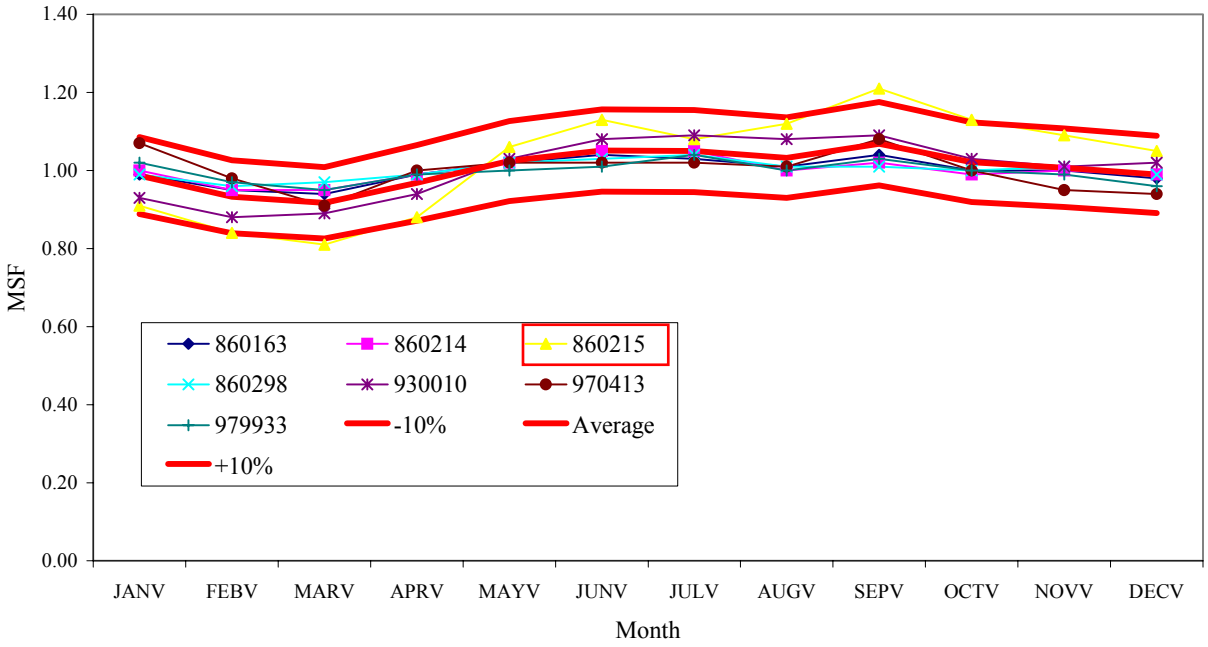


Figure 19. MSFs, Group Means, and ±10% Thresholds from the Group Mean for TTMSs in Factor Group 1

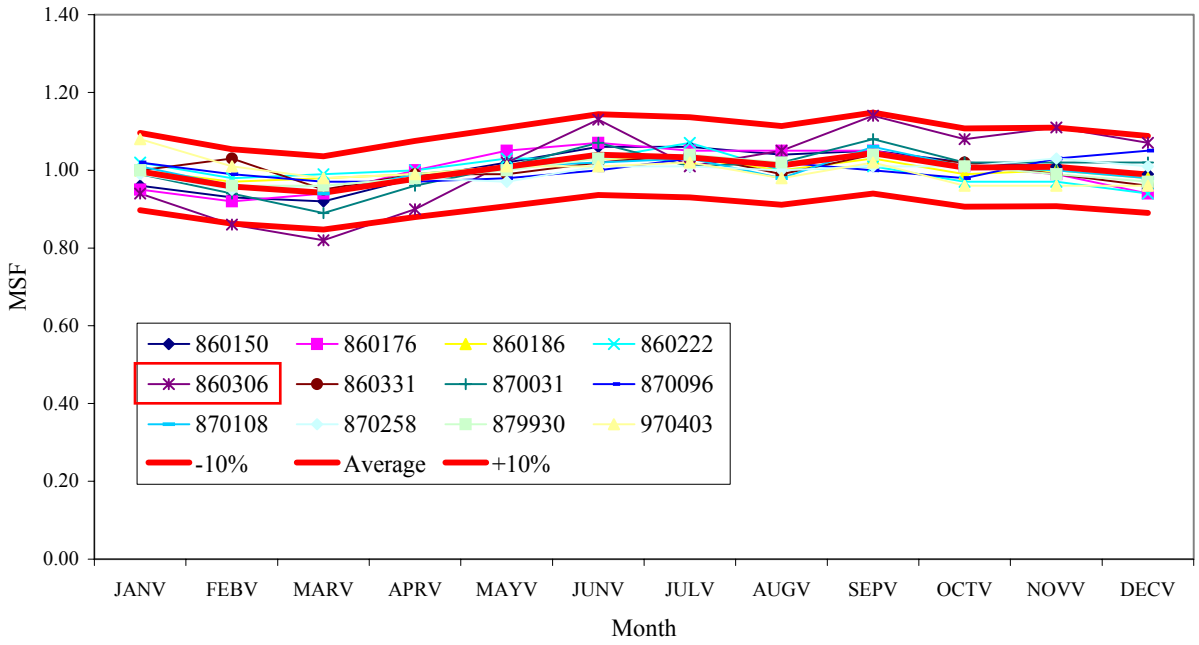


Figure 20. MSFs, Group Means, and ±10% Thresholds from the Group Mean for TTMSs in Factor Group 2

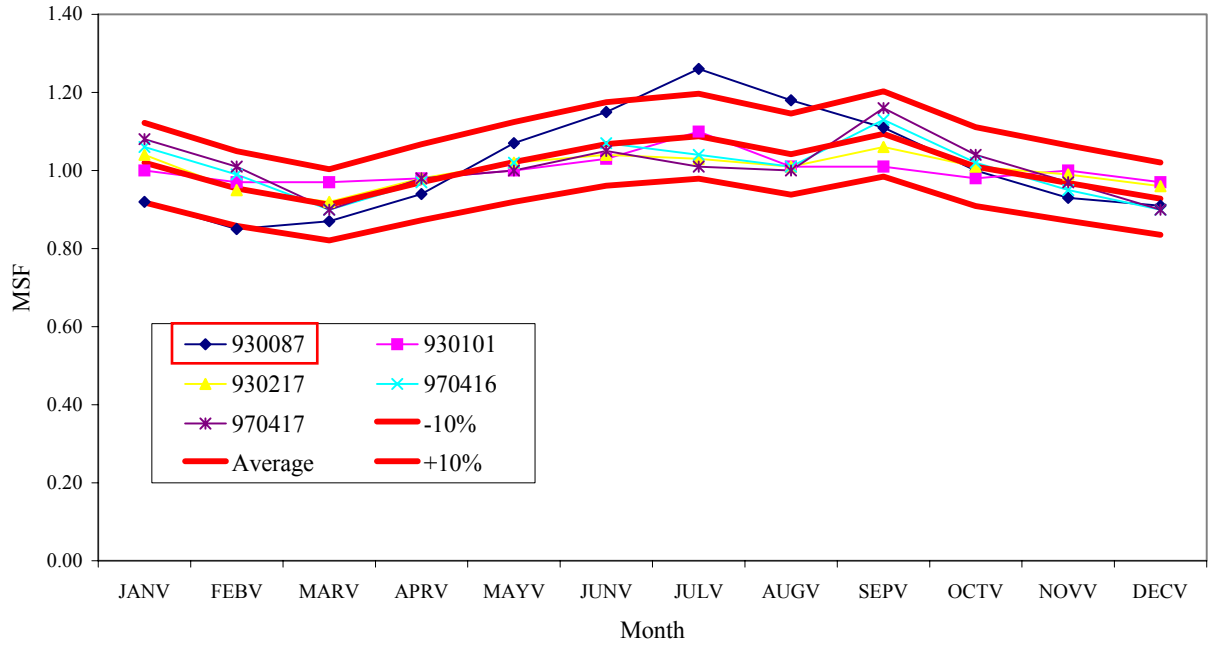


Figure 21. MSFs, Group Means, and ±10% Thresholds from the Group Mean for TTMSs in Factor Group 3

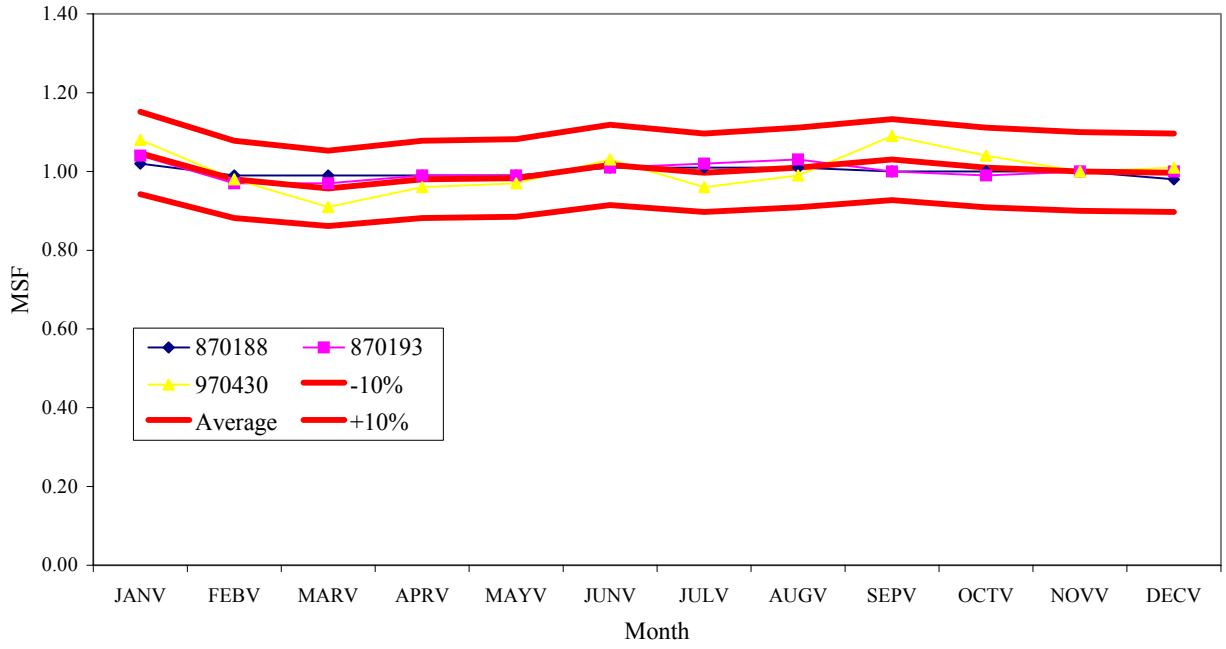


Figure 22. MSFs, Group Means, and ±10% Thresholds from the Group Mean for TTMSs in Factor Group 4

The results as shown in Figure 19 indicated that it was inappropriate to classify TTMS 860215 to Factor Group 1 since the MSFs for March and September fell outside the ±10% thresholds. Similarly, TTMSs 860306 and 930087 for Factor Groups 2 and 3, respectively, might not have been correctly classified. Consequently, these TTMSs were removed from the original factor groups determined by the model-based analysis. After examining their MSF patterns and locations, a new factor group, Group 5, was created for TTMSs 860215 and 860306 since the MSFs at these two count stations had similar seasonal fluctuations and they also were not far from each other (see Figure 18). The MSFs at TTMS 930087, on the other hand, suggested a much different seasonal variation and a unique factor group. Consequently, a new group, i.e., Group 6, was created for this isolated count station. TTMS 970430 of Group 4 was excluded from the remaining analysis due to a noticeably different MSF pattern observed at this location even though its MSFs were still within the thresholds of the group mean. Being the gateway to Key West and two national parks and surround by farm lands, this TTMS was situated at a unique location where there significant agricultural activities as well as tourist traffic, which was not typical of a location in an urban area.

As shown in Table 17 in Section 5.1.3, *SHP* (ratio of seasonal households to permanent households), *HMP3* (ratio of occupied hotel rooms to occupied hotel rooms plus households), *RETAILP* (retail workers as a percentage of total retail workers plus population), and *RH5_HQ* (percentage of retired householders of the highest income quartile) were identified as important factors in explaining the variations in the MSFs for the TTMSs in the tri-county area. They were subsequently used to further examine the four TTMSs that were manually classified or excluded. Table 33 shows the *SHP*, *HMP3*, *RETAILP*, and *RH5_HQ* values for the three manually classified TTMSs, i.e., 860215, 860306, and 930087.

Table 33. Land Use Variables at TTMSs 860215, 860306, and 930087

Group	TTMS	*SHP*	*HMP3*	*RETAILP*	*RH5_HQ*
5	860215	0.5392	0.2074	0.0900	0.1262
	860306	0.3051	0.3080	0.0540	0.0685
6	930087	0.5346	0.3897	0.7244	0.3414

As indicated in Table 33, the two TTMSs, i.e., 860215 and 860306, which were manually reclassified to form Group 5, shared similar land use and socioeconomic/demographic characteristics in addition to being in spatially proximity. Consequently, the reclassification appeared to be reasonable. Group 6 appeared to be also necessary based on the traffic and land use characteristics observed for TTMS 930087, since it had relatively higher *HMP3* and *RH5_HQ* than Group 5.

Table 34 shows the land use characteristics of the TTMSs that were originally classified to Group 4. A relatively higher *SHP*, *HMP3*, and *RH5_HQ* were observed at TTMS 970430. Similar to the profile of MSFs, these variables suggested a higher seasonal fluctuation for the traffic volumes observed at this specific TTMS. In other words, these statistics also supported the elimination of TTMS 970430 from Group 4.

Table 34. Land Use Variables at TTMSs 870188, 870193, and 970430

TTMS	*SHP*	*HMP3*	*RETAILP*	*RH5_HQ*
870188	0.0034	0.0000	0.0285	0.0528
870193	0.0085	0.0061	0.0851	0.0645
970430	0.0183	0.0129	0.0458	0.0156

After the questionable TTMSs were removed from the original groups, the mean MSFs for Groups 1 and 2 were nearly identical (see Figure 23). Since these two groups of TTMSs were adjacent to each other spatially, Group 1 and Group 2 were combined to form a new Group 1.

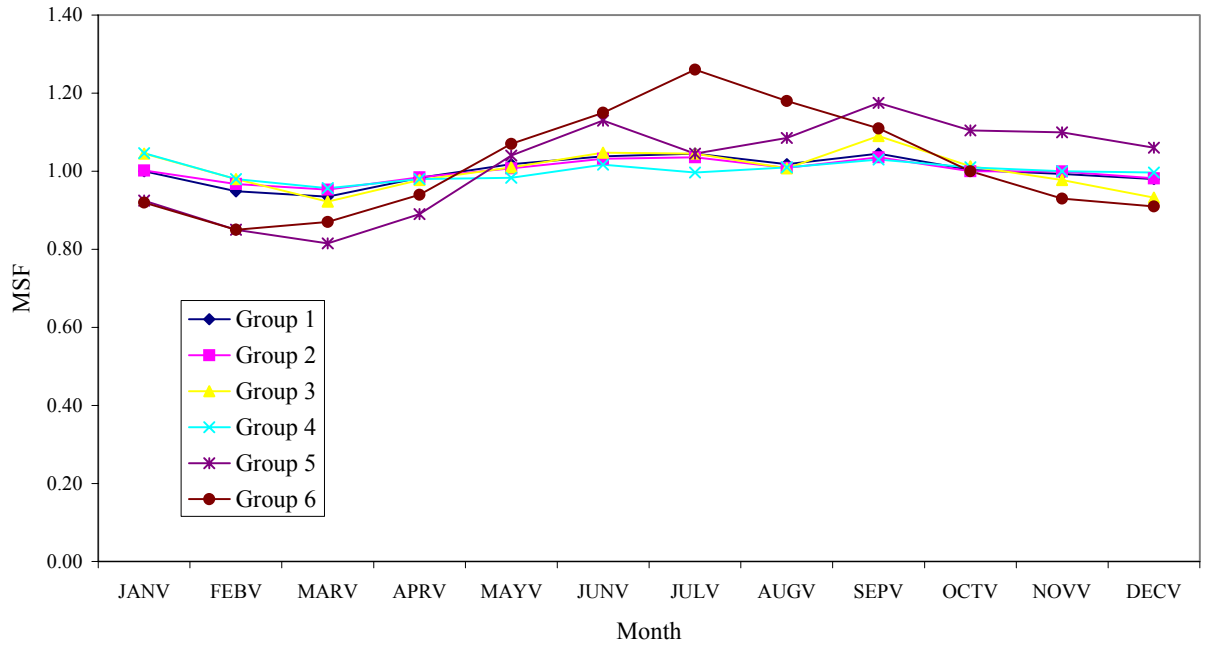


Figure 23. Intermediate MSF Group Means

There were five final groups categorized for the tri-county area as given in Table 35. Figure 24 shows the mean MSFs for these groups. These five finalized SF categories were subsequently assigned to the PTMSs in the tri-county area. The assignment procedure is described in the next section.

Table 35. Lookup Table for SF Groups in Tri-County Area

Grouping	Group Number					
Original	1	2	3	4	5	6
Final	1	1	2	3	4	5

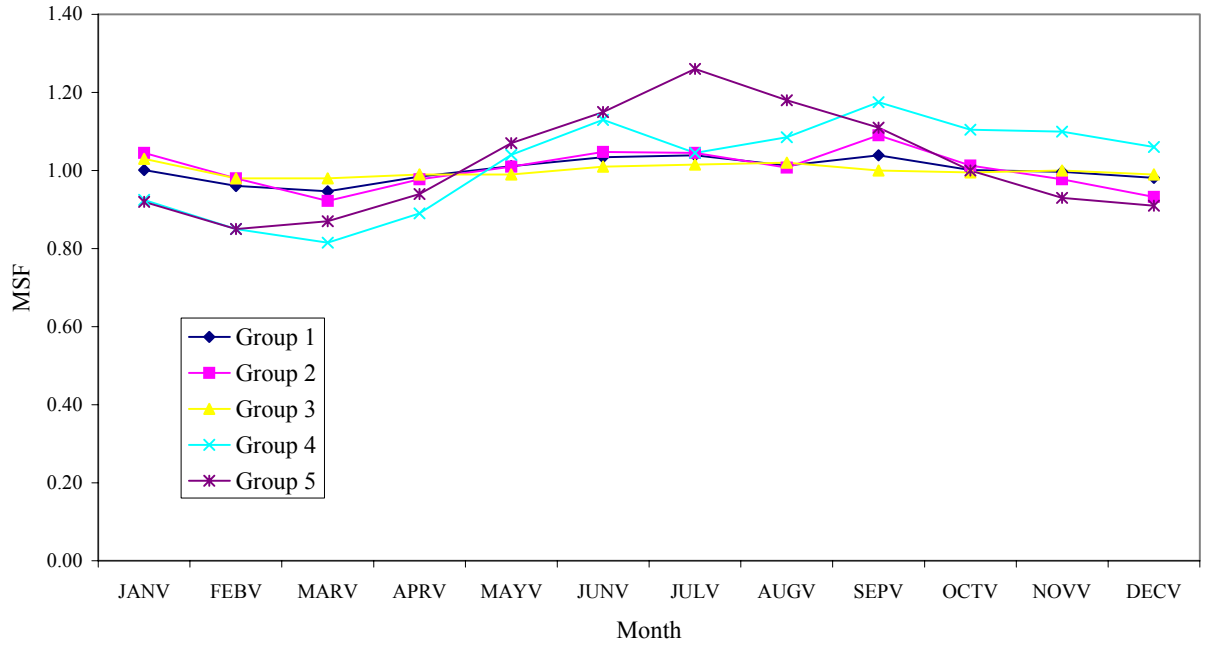


Figure 24. Five Final MSF Group Means

5.3.2 Seasonal Factor Assignment Procedure

A data-driven procedure was developed in this study to assign a seasonal factor category to a given PTMS. As mentioned in Section 5.1, for the urban areas in Southeast Florida, four significant factors that contributed to seasonal traffic patterns haven been identified and it has been suggested that the seasonal group of a PTMS could be determined based on its characteristics measured by these four variables. These four factors are:

1. Ratio of seasonal households to permanent households (*SHP*);
2. Hotel population or visitors (*HMP3*);
3. Ratio of retail employment to retail employment plus population (*RETAILP*); and
4. Percentage of retired households with high income (*RH5_HQ*).

Because the characteristics of any given PTMS may vary from those of the TTMSs, a method is needed to determine which seasonal group is the most appropriate for a given PTMS based on the four identified criteria. This was achieved through the application of fuzzy logic, an artificial intelligence technique, and a decision tree.

In fuzzy logic, membership in a given class may be partial. As an example, consider the following definitions for "short" people and "tall" people. Assume a person shorter than five feet is considered a "short" person and a person taller than six feet is a "tall" person. Applying these criteria, there is no difficulty in classifying a person into the "short person" class or the "tall person" class when this person has a height under five feet or above six feet. However, if the person's height falls in between five and six feet, then there is a "fuzziness" in the concept of "short" and "tall" and this person may be considered as both short and tall although the degree of "shortness" or "tallness" may vary depending on the person's height. In other words, such a

person may have a membership in both classes of "short person" or "tall person". If the person is closer to five feet, we may say this person's membership of the "short person" class is greater than that of the "tall person" class, and vice versa, and 5.5 feet may be considered as the break point where a person may be considered as equally short and tall.

Before assigning seasonal groups to PTMSs, the membership definitions for the five seasonal factor groups must be determined first based on the four land use variables. Once membership definitions (i.e., membership functions) are determined, a PTMS's membership of a particular seasonal group may be estimated with a probability (degree of belonging). Therefore, while a PTMS may have multiple memberships in several seasonal groups because it shares similar characteristics with a number of seasonal groups, its final membership or the seasonal group to be eventually assigned to it will depend on with which seasonal group the PTMS shares the most similarity. This is indicated by the highest probability value associated with its membership in this seasonal factor group. This fuzzy logic approach was implemented as a binary-split fuzzy decision tree, where the four land use variables, i.e., *SHP*, *HMP3*, *RETAILP*, and *RH5_HQ,* were sequentially considered to determine the membership of a PTMS. An overview for fuzzy decision tress is given in the next section, followed by a description of the assignment process that was developed to assign a specific SF category to each PTMS in the tri-county area.

5.3.2.1 *Overview of Fuzzy Decision Tree*

Decision trees are a specific decision analysis technique for analyzing various options or decisions for which risks and uncertainties exist. Using decision trees to reach a decision is effective since all choices may be examined, discussed, and challenged. Additionally, it helps make the best decisions on the basis of the existing information. A decision tree is commonly used to determine the classification of an instance. The input to construct a decision tree is known as a training dataset. The dataset contains records with several attributes, including a distinguished attribute called the class label. The goal of classification is to build a concise model in terms of the attributes from the records in the training dataset. The resulting model is then used to predict class labels for those without a class label.

Crisp decision tree techniques have been shown to be interpretable, efficient, problem-independent, and able to treat large-scale applications [OLA03]. However, they are also recognized as highly unstable classifiers with respect to minor perturbations in the data. The fuzzy sets formalism in fuzzy logic, instead of classifying an element being a member of a set or not, allows degrees of membership so that an element may simultaneously be a member of two or more fuzzy sets. Over the past few years, research has shown that fuzzy logic may introduce a promising improvement in enhancing stability and hence lead to better interpretability of decision tree induction [QUI86, JAN98, OLA03].

A fuzzy decision tree is a symbolic decision tree that incorporates the approximate reasoning provided by fuzzy representation [JAN98]. It combines the advantages of being popular in the real-world applications of learning from examples and high knowledge comprehensibility of decision tress as well as the ability of fuzzy representation to deal with inexact and uncertain information. As opposed to a classical decision tree, which gives only one class as the end

result, a fuzzy decision tree associates a set of probabilities for a given object with several or all classes. The following is a mathematical description of the fuzzy decision trees.

Assume that a sample set S consists of elements from mutually exclusive classes *P* and *N*. The probabilities for an object randomly selected from set S that belongs to *P* and *N* are $p/(p+n)$ and $n/(p+n)$, respectively, where *p* and *n* are the numbers of items for classes *P* and *N* in the sample set S. In Information Theory, a bit represents a binary digit and may assume a value of 0 and 1, with 0 representing one item and 1 the other. Therefore, a binary digit may be used to distinguish two items. Thus, *k* bits may have 2^k possible values and therefore may distinguish 2^k items. In other words, *n* items may be distinguished using $\log_2(n)$ bits. In addition, the optimal length of code to identify an object with probability *x* is $\log_2(1/x)$, i.e., $-\log_2(x)$ bits. For example, for an object *x* with a probability of 1/2, the optimal length code to identify *x* is $\log_2(1/2)$, i.e., 1. This is equivalent to distinguish two items by categorizing every *x* element into a set and all the elements other than *x* into the other set. The expected number of bits needed to encode the members that are randomly drawn from S for class *P* or *N* is known as entropy and is given as follows [QUI86]:

$$E(\mathrm{S}) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \tag{52}$$

Entropy measures the impurity of S. This is the information necessary to classify an example instance without any aids such as a decision tree. The class of an example instance is determined from scratch and no additional information is added to assist in the classification. The expected reduction in entropy due to introducing an attribute *A* to split the set S into subset S_1 and S_2, known as information gain, is defined as follows, where || represents the number of elements in the corresponding set:

$$Gain(\mathrm{S}, A) \equiv E(\mathrm{S}) - \sum_{k=1}^{2}\frac{|\mathrm{S}_k|}{|\mathrm{S}|}E(\mathrm{S}_k) \tag{53}$$

In classical set theory, a set may be defined by a two-valued characteristic function, i.e., $U \rightarrow \{0, 1\}$ where *U* is the universe of discourse. In fuzzy set theory, however, a fuzzy subset of the universe of discourse *U* is described by a membership function, $\mu_v(V)$: $U \rightarrow [0,1]$, which represents the degree to which $\mu \in U$ belongs to the set *v*. A fuzzy linguistic variable is an attribute whose domain contains linguistic values known as fuzzy terms, which are labels for the fuzzy sestets. For example, consider the continuous attribute *SHP* of the TTMSs located in the tri-county area. This demographic attribute becomes a fuzzy linguistic variable when two linguistic terms, e.g., low and high, are used as domain values and there exists an overlap (i.e., fuzziness) between these two terms.

Figure 25 illustrates the associated fuzzy sets and probable memberships for the *SHP* fuzzy variable. The figure shows that when the *SHP* value for a given TTMS is less than SHP_1, a count station is considered to have a low *SHP* value. When the *SHP* value is larger than SHP_2, a count station is considered to have a high *SHP* value. However, when the *SHP* value falls

between SHP_1 and SHP_2, the count station will have two membership probabilities for the low and high classes, respectively, depending on where the *SHP* value fell.

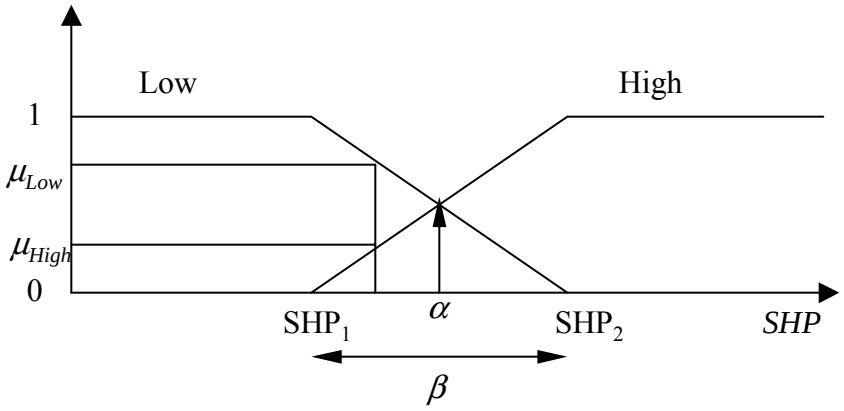


Figure 25. Fuzzy Subsets and Memberships for *SHP* Attribute

Fuzzy sets are generally described with convex functions peaking at 1. The two parameters defining the membership function are: α, which is the location of the cut-point (or the break point) and corresponds to the split threshold, and β, which is the degree of spread, i.e., width, that defines the transition region on the attribute [OLA03]. For a given *SHP* that is less than $\alpha - \beta/2$ or larger than $\alpha + \beta/2$, the membership function is 1. In other words, the object is 100% classified to the low or high fuzzy subsets. On the other hand, a given *SHP* value may be classified to both the low and high subsets with different degrees, i.e., μ_{Low} and μ_{High}, where the sum of memberships is equal to 1. Piecewise linear functions similar to the one illustrated in Figure 25 are the most widely used discriminator functions to fuzzily split an observation into fuzzy sets. This membership function would fuzzily partition a fuzzy variable such as *SHP* into two overlapping subsets, i.e., low and high.

Figure 26 shows the split of a tree node corresponding to a fuzzy set S into two fuzzy sestets S_1 and S_2 based on the chosen attribute *A* at the node S where a_i is the *i*th item in the data set; $A_1(a_i)$ and $A_2(a_i)$ denote the memberships for item a_i for fuzzy subsets S_1 and S_2, respectively; and μ_S, μ_{S_1}, and μ_{S_2} denote the cumulative memberships for fuzzy sets S, S_1 and S_2, respectively. Note that $\mu_S(a_i)$ is equal to 1 when set S is at the root and that the sum of $A_1(a_i)$ and $A_2(a_i)$ is equal to 1.

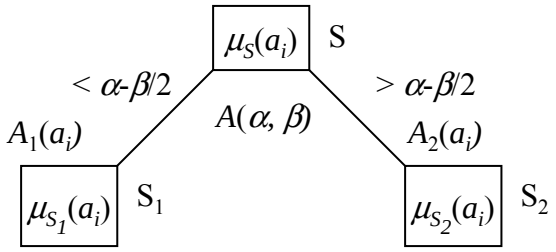


Figure 26. Node Partition in a Fuzzy Decision Tree

The cumulated memberships for fuzzy subsets S_1 and S_2 are calculated as follows:

$$\mu_{S_1}(a_i) = \mu_S(a_i) \times A_1(a_i) \tag{54}$$

$$\mu_{S_2}(a_i) = \mu_S(a_i) \times A_2(a_i) \tag{55}$$

In the following section, a procedure developed to create a fuzzy decision tree for assigning a specific SF category to a PTMS is described.

5.3.2.2 *Construction of Fuzzy Decision Tree*

The fundamental concept of utilizing the fuzzy decision tree technique is to select each split of a subset so that the data in each of the descendant subsets are composed of less SF categories in the parent subset. The purpose is to develop an objective classification rule that may determine the category of a given PTMS. The steps to create a fuzzy decision tree for assigning the five SF categories (or classes), i.e., c_1, …, c_5, to PTMSs are as follows:

Step 1. Sort the TTMSs by the values of each of the fuzzy variables to generate sequences of ordered values, a_1, …, a_N; where N is the number of TTMSs in the study area;

Step 2. For each attribute a, calculate the average distance of data points as the overlapping width as follows:

$$\beta = \frac{1}{N-1}\sum_{i=1}^{N-1} a_{i+1} - a_i \; ; \tag{56}$$

Step 3. Select a specific attribute a, e.g., *SHP*, for the TTMSs in the study area as the current fuzzy variable, and denote the sample set as S;

Step 4. For each data point between $a_1 + \beta/2$ and $a_N - \beta/2$, generate candidate cut point as the average of two adjacent data points, i.e.,

$$\alpha = \frac{a_i + a_{i+1}}{2}; \tag{57}$$

Step 5. For each α, calculate the information gain, i.e., $E_F(S) - E_F(A, \alpha, S)$, according to the following equations:

$$E_F(\mathrm{S}) = -\sum_{j=1}^{5} p(c_j, \mathrm{S}) \times \log_2 p(c_j, \mathrm{S}); \tag{58}$$

$$p(c_j, \mathrm{S}) = \sum\nolimits_{a_i \in c_j} \mu_{\mathrm{S}}(a_i) \Big/ \sum\nolimits_{a_i \in \mathrm{S}} \mu_{\mathrm{S}}(a_i); \tag{59}$$

$$E_F(A, \alpha; S) = \frac{N_F^{S_1}}{N_F^S} \times E_F(\mathrm{S}_1) + \frac{N_F^{S_2}}{N_F^S} \times E_F(\mathrm{S}_2); \tag{60}$$

$$E_F(\mathrm{S}_1) = -\sum_{j=1}^{5} p(c_j, \mathrm{S}_1) \times \log_2 p(c_j, \mathrm{S}_1); \quad (61)$$

$$E_F(\mathrm{S}_2) = -\sum_{j=1}^{5} p(c_j, \mathrm{S}_2) \times \log_2 p(c_j, \mathrm{S}_2); \quad (62)$$

$$p(c_j, \mathrm{S}_k) = \frac{N_F^{\mathrm{S}_k c_j}}{N_F^{\mathrm{S}_k}}, \; k = 1, 2; \quad (63)$$

$$N_F^{\mathrm{S}} = \sum_{i=1}^{|\mathrm{S}|} \mu_S(a_i); \quad (64)$$

$$N_F^{\mathrm{S}_1} = \sum_{i=1}^{|\mathrm{S}|} \mu_{\mathrm{S}_1}(a_i); \quad (65)$$

$$N_F^{\mathrm{S}_2} = \sum_{i=1}^{|\mathrm{S}|} \mu_{\mathrm{S}_2}(a_i); \quad (66)$$

$$N_F^{\mathrm{S}_k c_j} = \sum_{a_i \in c_j} A_k(a_i), \; k = 1, 2; \quad (67)$$

where

$E_F(\mathrm{S})$ = fuzzy class entropy in S;

$E_F(A, \alpha, \mathrm{S})$ = class information entropy for attribute A in respect to a given α calculated with the probability of fuzzy partition, i.e., the proportion of examples in S that belongs to class c_j;

$E_F(\mathrm{S}_k)$ = fuzzy class entropy in S_k, $k = 1, 2$;

$p(c_j, \mathrm{S})$ = fuzzy proportion of examples in S;

$p(c_j, \mathrm{S}_k)$ = fuzzy proportion of examples in S_k, $k = 1, 2$;

A_1, A_2 = membership functions for fuzzy sets S_1 and S_2;

c_j = the jth SF category;

j = SF category number, $j \in 1, \ldots, 5$;

N = the total number of TTMSs;

$N_F^{\mathrm{S}_k c_j}$ = sum of the memberships for elements belong to class c_j in fuzzy set S_k in S_k, $k = 1, 2$;

N_F^{S} = sum of the memberships for elements in fuzzy set S;

$N_F^{\mathrm{S}_k}$ = sum of the memberships for elements in fuzzy set S_k, $k = 1, 2$; and

S_1, S_2 = fuzzy sets (tree branches) 1 and 2.

Step 6. Select α that gives the maximum information gain; repeat Steps 1 to 5 for other fuzzy variables;

Step 7. Select the variable, which produces the maximum information gain, to generate two child branches and nodes;

Step 8. Calculate the truth level for each branch as follows:

$$\eta_1 = \frac{N_F^{S_1}}{N_F^S},\ \eta_2 = \frac{N_F^{S_2}}{N_F^S} \tag{68}$$

If $\eta_1 \le \lambda$ or $\eta_1 \le \lambda$, where λ is 0.1, delete the corresponding branches. Otherwise, calculate the truth level of each branch belonging to the *j*th class as follows:

$$\mu_{1j} = \frac{\sum_{a_i \in c_j} A_1(a_i)}{N_F^{S_1}},\ \mu_{2j} = \frac{\sum_{a_i \in c_j} A_2(a_i)}{N_F^{S_2}} \tag{69}$$

If $\max_{j=1}^{k}(\mu_{1j}) \ge \gamma$ or $\max_{j=1}^{k}(\mu_{2j}) \ge \gamma$, where γ is 0.9, terminate the corresponding branch as a leaf and assign this leaf as the class c_j.

Step 9. Repeat Steps 3 to 8 to create additional branches.

As mentioned previously, four fuzzy attributes were considered in the construction of fuzzy decision tree. They were *SHP*, *HMP3*, *RETAILP*, and *RH5_HQ*. In Step 1, the values of fuzzy variables are sorted to produce *N* ordered sequences. The average distance of data points was then used as the overlapping width (β) for each of the fuzzy variable. Generally speaking, wide overlaps mean high uncertainty. Since the true nature of fuzziness for a given attribute is unknown, the data obtained from the field was utilized to heuristically estimate β [PEN01]. The method for approximating the width parameter was assumed to incorporate the characteristics of uncertainty of the associated fuzzy variable in the process of modeling membership functions. As a result, the β value associated with each fuzzy variable remained a constant during the construction of fuzzy decision tree.

In Step 4, every pair of adjacent data points, i.e., a_i and a_{i+1}, within the interval of $[a_1 + \beta/2, a_N - \beta/2]$ suggested a potential partition to create a cut point (α). This was because a cut point between a_i and a_{i+1} would not lead to a partition that had the maximum information gain in classification if these two data points belonged to the same class [PEN01]. The resulted partition allowed at least one data point to completely classify to each fuzzy subset. This was to ensure a valid classification based on the observed values for a given fuzzy variable.

In Step 5, the corresponding information gain for a candidate α was calculated for the current fuzzy variable. In Step 6, the α value that gave the maximum information gain for a specific fuzzy variable was identified and the process moved to determining α for the other fuzzy variables by repeating Steps 1 to 5. The attribute that produced the maximum information gain was then selected as the variable to fuzzily split set S into sets S_1 and S_2 in Step 7.

In Step 8, the branches split from the attribute that yielded the maximum information gain was evaluated to determine if the sum of the memberships for every element that were classified to a given branch, i.e., η_k, where $k \in \{1, 2\}$, was significant enough (i.e., $> \lambda$). If not, the corresponding branch was eliminated to assure the simplicity of the decision tree. Otherwise, the sum of the memberships for the elements that belonged to a given class on each branch, i.e., μ_{kj}, where $k \in \{1, 2\}$ and $j \in \{1, 2, 3, 4, 5\}$, was then calculated. If the maximum was greater than γ,

the tree was terminated at this node and no additional split was processed. Otherwise, repeat Steps 3 through 8 until the stopping criterion was met..

5.3.2.3 *SF Category Assignment*

After the decision tree was constructed, the membership of a given entity to each decision node was calculated. Figure 27 shows the conceptual fuzzy decision tree constructed using the five SF categories and the four fuzzy attributes of the 26 TTMSs in the tri-county area. The leaf nodes of the tree indicate the seasonal groups that have been classified. The tree describes how a seasonal group may be determined based on the values of the four land use variables. For instance, SF category 4 was characterized by a higher value of hotel population but a lower value of retail employment. SF category 3 was associated with a smaller hotel population, a lower seasonal household percentage, and a higher percentage of high-income retired households. However, SF category 2 could not be distinguished from SF category 1 using the decision tree illustrated in Figure 27. In this case, some engineering judgment must be exercised to determine whether a count station belongs to SF category 1 or 2. Based on the results of modal-based cluster analysis results, it may be reasonable to assign a count station to SF category 1 or 2 depending on to which group it is closer spatially. Note that the low and high values for the same fuzzy attribute at different nodes were different because they were determined based on the values of the attributes of the remaining samples that were to be classified.

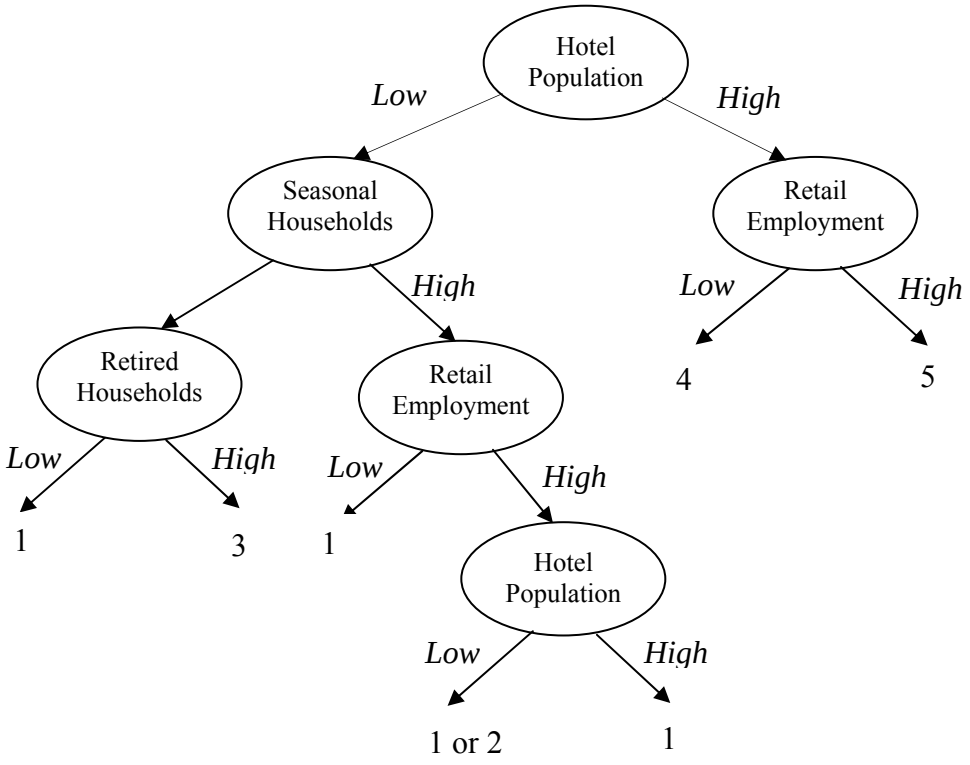


Figure 27. Conceptual Fuzzy Tree for Classification of TTMS Groups

Figure 28 shows the fuzzy decision tree in more detail with α and β calculated for each node and the number of samples that were fully or fuzzily classified to the low and high sets. The figure shows that *HMP3* yielded the maximum information gain when α was equal to 0.1384. Consequently, this attribute was placed at the root node, HMP3_1. Two branches, HMP3_1_L and HMP3_1_R, were grown from the root. At the HMP3_1 node, no TTMSs had a HMP3 value that fell within the fuzzy interval, i.e., F (fuzziness) is zero. According to the membership function derived in the process, 23 out of the 26 TTMSs were 100% classified to HMP3_1_L and the remaining three TTMSs were classified to HMP3_1_R. In other words, the HMP3 attribute was low at 23 TTMSs and high at three TTMSs.

The second attribute identified in the process was *SHP* when α was equal to 0.0313. Five TTMSs were partially partitioned into the SHP_2_L and SHP_2_R branches at the SHP_2 node according to the estimated memberships, i.e., F was equal to 5. In total, eight and 20 TTMSs were classified to the SHP_2_L and SHP_2_R branches, respectively. The TTMSs on the SHP_2_L branch could be interpreted as those with lower *HMP3* and lower *SHP*. Similarly, lower *HMP3* and higher *SHP* were observed of the TTMSs on SHP_2_R.

The third attribute identified in the process was *HQ*, a short form of *RH5_HQ*, when α was equal to 0.0485. At the HQ_3 node, five TTMSs from SF category 1 and one TTMS from SF category 4 were 100% classified to the HQ_3_L and HQ_3_R branches, respectively. Additionally, the process identified two TTMSs with their *RH5_HQ* values in the overlapping area of the low and high subsets. The figure shows that the decision tree did not grow beyond the HQ_3 node and two classes, i.e., 1 and 3, were identified. The results showed that it might be appropriate to assign SF category 1 to PTMSs with low *HMP3*, *SHP*, and *RH5_HQ* attribute values. On the other hand, SF category 3 may be appropriate for PTMSs with similar *HMP3* and *SHP* values but a high *RH5_HQ* value.

The retail attribute was the fourth fuzzy variable included in the decision tree. When α was equal to 0.0707, four TTMSs in SF category 1 and ten TTMSs in both categories 1 and 3 were completely classified to Retail_4_L and Retail_4_R, respectively, at the Retail_4 node. Six TTMSs were fuzzily partitioned.

The *HMP3* attribute was selected again in the process as the fifth fuzzy variable. When α was equal to 0.0397, ten TTMSs from either SF category 1 or 3 were completely classified to HMP3_5_L. No additional information gain could be achieved by further classifying the tree beyond this node and the growing of the decision tree was consequently terminated. Figure 28 also shows that three TTMSs were classified to HMP3_5_R while three TTMSs were fuzzy. Furthermore, the last attribute identified in the process was the retail attribute. When α was equal to 0.4072, TTMSs from SF categories 4 and 5 were explicitly partitioned into two separate branches.

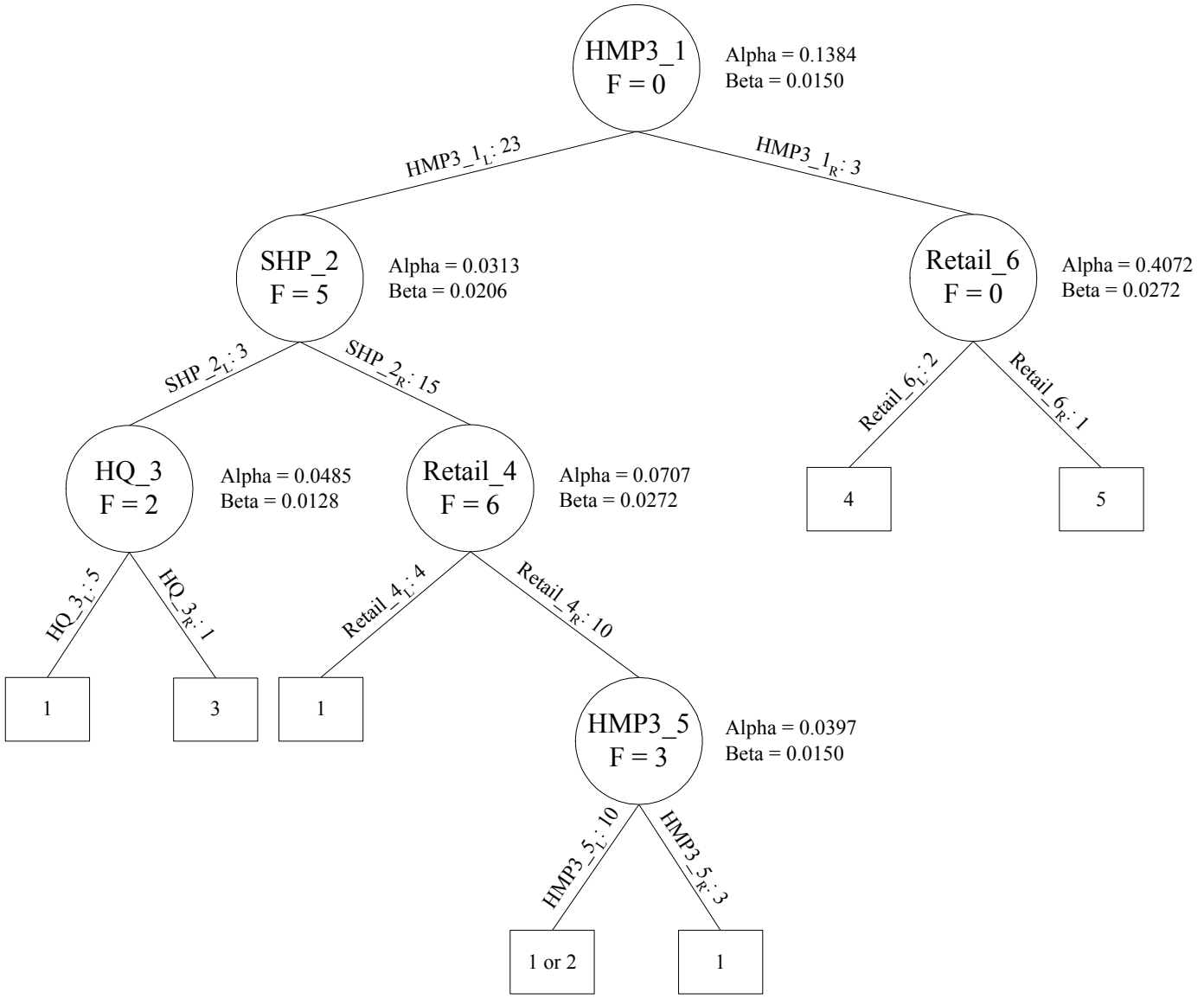


Figure 28. Fuzzy Decision Tree for Assigning SF Categories

Tables 36 through 40 give the results of classifications at nodes 1 through 5 (HMP3_1, SHP_2, HQ_3, RETAIL_4, and HMP3_5). Table 36 lists the land use characteristics of the TTMSs in categories 4 and 5 after they were classified to the right branch of the HMP3_1 node. It may be seen that Category 5 was characterized by a high percentage of hotel/motel population, seasonal households, high-income retired households, and high ratio of retail employment to retail employment plus population. However, *RETAIL* was the variable that distinguished category 4 from category 5.

Table 36. Land Use Characteristics of TTMSs in Categories 4 and 5 Sorted by *HMP3*

Site	Cluster	*HMP3*	*RETAIL*	*SHP*	*RHP5_HQ*
930087	5	38.97	0.7244	53.46	34.14
860306	4	30.80	0.0540	30.51	6.85
860215	4	20.74	0.0900	53.92	12.62

All the TTMSs in categories 1, 2, and 3 were 100% classified to the left branch of the HMP3_1 node because they had much lower hotel/motel population. Next, they were split again at the SHP_2 node based on the seasonal household percentages. Table 37 gives the TTMSs in this group, sorted by their *SHP* values. The 15 TTMSs above the shaded area were considered to have a high percentage of seasonal households and were classified to the right branch of the SHP_2 node. The four TTMSs below the shaded area were considered to have a low percentage of seasonal households and were classified to the left branch of the SHP_2 node. The five TTMSs in the shaded area fell in the fuzzy range and had partial memberships in both the high and low value groups.

Table 38 shows how the four TTMSs that were fully classified and the five TTMSs that were partially classified to the left branch of the SHP_2 node were classified into left and right braches of the HQ_3 node. One TTMS above the shaded area was considered to have a high percentage of high-income retired households; four had a low percentage; and three in the shaded area had partial memberships in both the high and low groups.

Tables 39 and 40 illustrate how the TTMSs were classified into high and low groups at the RETAIL_4 and HMP3_5 nodes in a similar manner. Again, the TTMSs above the shaded area belonged to the high value group, those below the shaded area belonged to the low value group, and those in the shaded area had partial memberships in both.

Keep in mind that the sequence of the variables that were used to classify the TTMSs was not determined based on which remaining attribute presented largest value differences, but was obtained by maximizing information gain at each step in the process.

Table 37. Land Use Characteristics of TTMSs in Categories 1, 2, and 3 after the SHP_2 Node

Site	Cluster	*HMP3*	*SHP*	*RETAIL*	*RHP5_HQ*
860150	1	4.10	34.1292	0.1565	1.4351
860214	1	4.58	26.2125	0.0506	6.3984
860176	1	5.95	16.9673	0.1023	1.9245
970413	1	0.00	15.4343	0.0423	13.3286
930217	2	2.70	14.0361	0.1090	7.1229
930010	1	3.84	14.0116	0.1425	6.2669
860331	1	3.36	13.1698	0.0823	3.8843
860163	1	2.65	11.3742	0.0967	5.5310
970417	2	0.74	10.0141	0.0860	5.8071
970416	2	3.09	9.1122	0.1126	6.4103
870031	1	2.14	8.1126	0.0512	6.7755
870108	1	6.02	6.6136	0.0821	2.7479
930101	2	0.69	4.9795	0.1565	2.6759
970403	1	0.89	4.4534	0.0634	2.3952
860186	1	1.44	4.2436	0.0727	2.3015
870258	1	0.80	3.9788	0.0176	0.9370
879930	1	0.14	3.7500	0.0687	1.5245
860298	1	0.97	3.4925	0.1228	1.4752
979933	1	0.79	3.2608	0.0800	3.0345
860222	1	6.94	2.9888	0.1720	4.4286
870096	1	0.75	1.0011	0.0454	2.4563
870193	3	0.61	0.8521	0.0851	6.4546
870188	3	0.00	0.3386	0.0285	5.2755

Table 38. Land Use Characteristics of TTMSs in Categories 1 and 2 after the HQ_3 Node

Site	Cluster	*RHP5_HQ*	*HMP3*	*SHP*	*RETAIL*
870193	3	6.4546	0.6100	0.8521	0.0851
870188	3	5.2755	0.0000	0.3386	0.0285
860222	1	4.4286	6.9400	2.9888	0.1720
979933	1	3.0345	0.7900	3.2608	0.0800
870096	1	2.4563	0.7500	1.0011	0.0454
879930	1	1.5245	0.1400	3.7500	0.0687
860298	1	1.4752	0.9700	3.4925	0.1228
870258	1	0.9370	0.8000	3.9788	0.0176

Table 39. Land Use Characteristics of TTMSs in Categories 1 and 2 after the Retail_4 Node

Site	Cluster	*RETAIL*	*HMP3*
860222	1	0.1720	6.9400
860150	1	0.1565	4.1000
930101	2	0.1565	0.6900
930010	1	0.1425	3.8400
860298	1	0.1228	0.9700
970416	2	0.1126	3.0900
930217	2	0.1090	2.7000
860176	1	0.1023	5.9500
860163	1	0.0967	2.6500
970417	2	0.0860	0.7400
860331	1	0.0823	3.3600
870108	1	0.0821	6.0200
979933	1	0.0800	0.7900
860186	1	0.0727	1.4400
879930	1	0.0687	0.1400
970403	1	0.0634	0.8900
870031	1	0.0512	2.1400
860214	1	0.0506	4.5800
970413	1	0.0423	0.0000
870258	1	0.0176	0.8000

Table 40. Land Use Characteristics of TTMSs in Categories 1 and 2 after the HMP3_5 Node

Site	Cluster	*HMP3*	*SHP*	*RETAIL*	*RHP5_HQ*
860222	1	6.9400	2.9888	0.1720	4.4286
870108	1	6.0200	6.6136	0.0821	2.7479
860176	1	5.9500	16.9673	0.1023	1.9245
860150	1	4.1000	34.1292	0.1565	1.4351
930010	1	3.8400	14.0116	0.1425	6.2669
860331	1	3.3600	13.1698	0.0823	3.8843
970416	2	3.0900	9.1122	0.1126	6.4103
930217	2	2.7000	14.0361	0.1090	7.1229
860163	1	2.6500	11.3742	0.0967	5.5310
860186	1	1.4400	4.2436	0.0727	2.3015
860298	1	0.9700	3.4925	0.1228	1.4752
970403	1	0.8900	4.4534	0.0634	2.3952
979933	1	0.7900	3.2608	0.0800	3.0345
970417	2	0.7400	10.0141	0.0860	5.8071
930101	2	0.6900	4.9795	0.1565	2.6759
879930	1	0.1400	3.7500	0.0687	1.5245

Table 41 shows the final memberships of the 26 TTMSs included in the development of fuzzy decision tree on the end braches of the decision tree, i.e., HQ_3_L, HQ_3_R, $Retail\_4_L$, $HMP3\_5_L$, $HMP3\_5_R$, $Retail\_6_L$, and $Retail\_6_R$. The table shows that TTMS 860222 from SF category 1 and TTMS 870188 from SF category 4 were the TTMSs that were fuzzily partitioned since the memberships in both of the HQ_3_L and HQ_3_R columns were non-zero. The HQ_3_R membership of TTMS 860222 was not significant, i.e., the land-use attributes observed for TTMS 860222 resembled more closely those on HQ_3_L than those on HQ_3_R. Similarly, TTMS 870188 was more likely to be classified to HQ_3_R. In other words, the TTMSs could be explicitly spilt into two subsets, i.e., SF categories 1 and 4, after three land-use attributes were incorporated in the decision tree. The results suggested assigning SF category 1 to the PTMSs with lower *HMP3*, *SHP*, and *RH5_HQ* values. For the PTMSs with similar *HMP3* and *SHP* but higher *RH5_HQ* values, SF category 4 might be a better choice.

The fuzzy decision tree presented in Figure 28 was subsequently applied to determine the SF categories for the PTMSs in the tri-county area. For the PTMSs that were partitioned to multiple terminated nodes in the decision tree, the seasonal category with the highest membership was selected. For those that were classified to $HMP3\_5_L$, the SF category was determined by considering the distance between the TTMS group and the location of the PTMSs. Figure 29 shows the assignment results.

Table 41. Fuzzy Decision Tree Memberships for 26 TTMSs in the Tri-County Area

SF Category	Site	HQ_3$_L$	HQ_3$_R$	Retail_4$_L$	HMP3_5$_L$	HMP3_5$_R$	Retail_6$_L$	Retail_6$_R$	Sum
1	860150				0.4133	0.5867			1
	860163				1				1
	860176					1			1
	860186			0.4265	0.5735				1
	860214			1					1
	860222	0.4726	0.0954			0.432			1
	860298	0.3252			0.6748				1
	860331			0.0735	0.8401	0.0864			1
	870031			1					1
	870096	1							1
	870108			0.0809		0.9191			1
	870258	0.0874		0.9126					1
	879930	0.1990		0.4594	0.3416				1
	930010				0.5867	0.4133			1
	970403			0.7684	0.2316				1
	970413			1					1
	979933	0.4369		0.0890	0.4741				1
2	930101				1				1
	930217				1				1
	970416				1				1
	970417				1				1
3	870188	0.1680	0.8320						1
	870193		1						1
4	860215						1		1
	860306						1		1
5	930087							1	1

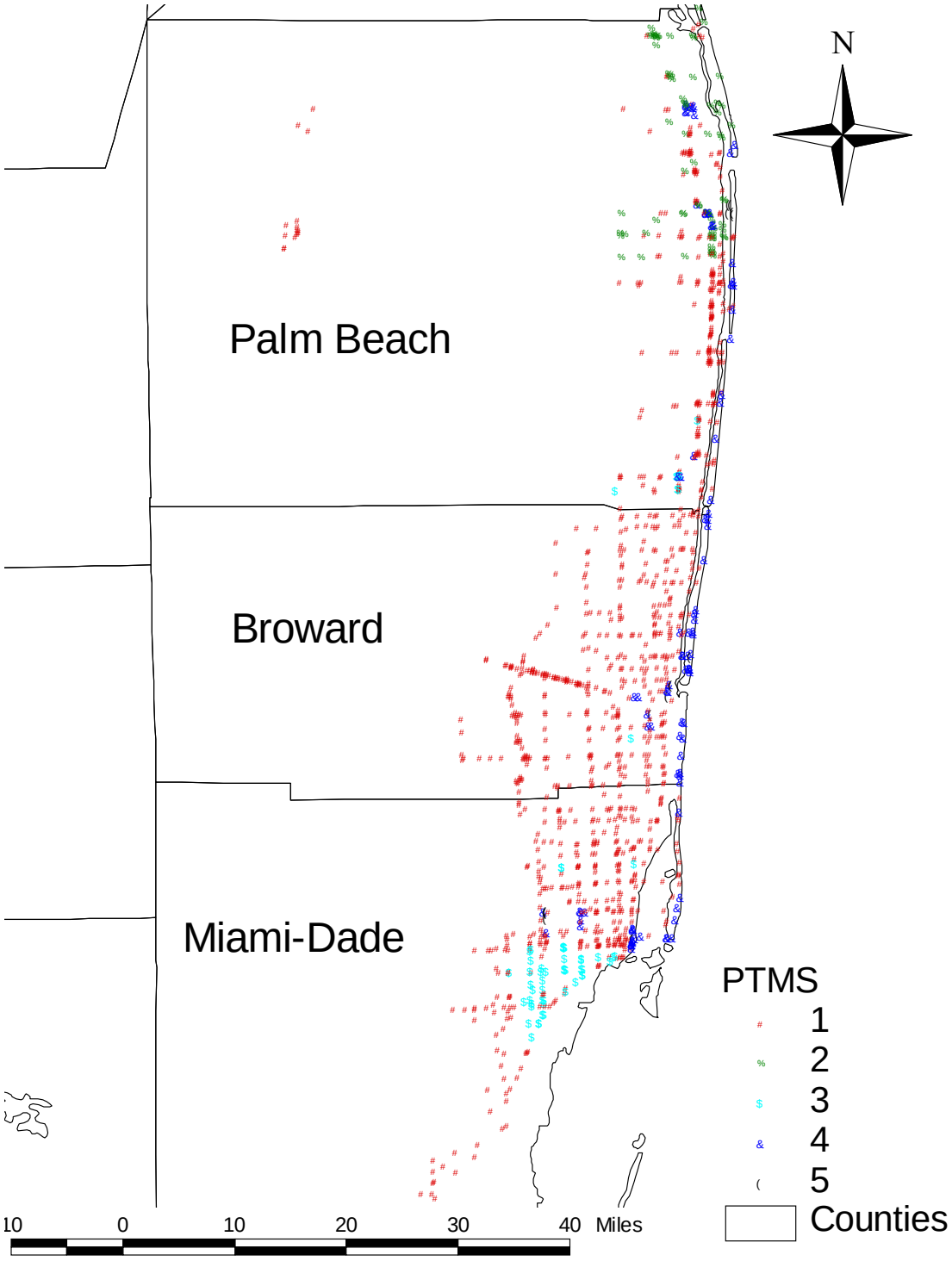


Figure 29. SF Category Assignment Result

6. A PROTOTYPE GIS APPLICATION

A prototype software program, FloridaSFAP (Florida Seasonal Factor Analysis Program) was developed to support seasonal factor analyses. The program was implemented with a GIS user interface. It supports visualization of land use, demographic, socioeconomic, transportation, and traffic data, as well as cluster and regression analyses with SAS and S-Plus programs. The program combines various data sources that are commonly available to transportation professionals. The user may retrieve data from the Census 2000, FDOT Roadway Characteristics Inventory (RCI) database, Traffic Information CD, and the employment data purchased by FDOT. For demographic and socioeconomic data, five geographic structures are available in the program: county, metropolitan planning organization (MPO) jurisdiction, census block group, census tract, and traffic analysis zone. Census data at block group level are available for all counties, and include population, permanent and seasonal households, income groups, age groups, income groups, etc. TAZ data are only available for urban areas that have established Florida Standard Urban Traffic Model Structure (FSUTMS) models, and include households, population, number of occupied hotel/motel rooms, industrial employment, commercial employment, service employment, total employment, and school enrollment at TAZ level. These data are typically either compiled based on census data or are estimated by county planning departments or MPOs. The data developed for the 1999 transportation models for Broward, Miami-Dade, and Palm Beach counties were included in the software developed for this research. The employment database reflects the 2001 employment and covers the entire Florida.

The program was developed within ArcView®, an Environmental System Research Institute product, and was customized with Avenue, an ArcView script language, and VisualBasic®. To allow people with limited knowledge of ArcView or GIS to use the program, it was designed as a menu-driven program. Several customized tools were also developed to allow the user to interact with the statistical analysis programs for SAS and S-plus. The top-level graphic user interface in FloridaSFAP is illustrated in Figure 30. The menus to the right of the *Window* menu are customized menus, which are not part of the standard ArcView menu. The functions provided by the customized menus are:

- Study Area – define study area, select spatial analysis units, and specify folder for the TTMS and PTMS data from the FDOT Traffic Information CD.
- Count Stations – display for selected TTMSs and PTMSs their site identifiers, seasonal factor groups, functional classes, traffic volumes, number of lanes, buffer size, and the values of the four land use variables.
- Seasonal Groups – allows the user to examining and modify the existing seasonal factor groups. Queries may be made regarding seasonal traffic profiles and land use variables of selected TTMSs and seasonal factor group statistics.
- Land Use – display maps that show the aerial photograph and demographic and socioeconomic data distribution in the study area;
- Network – display the transportation network with functional classification in the study area.

- Advanced – provide the user with access to customized tools including buffer analysis, regression analysis, hierarchical cluster analysis, and model-based cluster analysis, and display the results.
- Clear – clear the map area of previous display results.

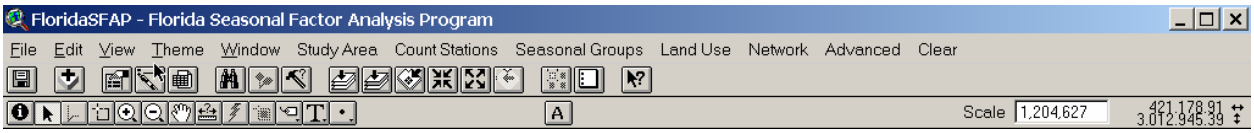


Figure 30. Top-Level Menu in FloridaSFAP

In the following sections, the design of the user interface and the functions of the program are described.

6.1 Study Area Menu

The *Study Area* menu, as shown in Figure 31, provides the user the options to select a study area previously defined or to create a new one for which seasonal factor analysis is desired. Selecting "Existing Study Area" will cause a dialog box shown in Figure 32 to be displayed for the user to select a previously defined study area.

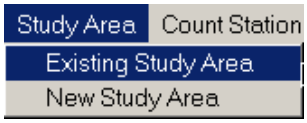


Figure 31. *Study Area* Menu

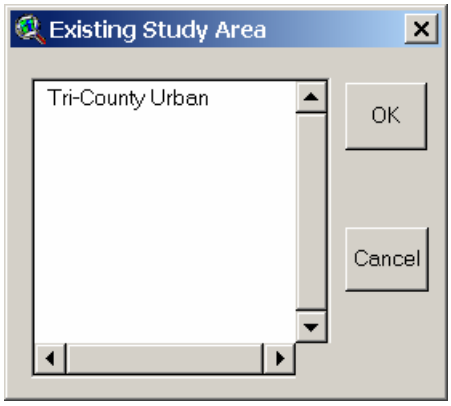


Figure 32. Selection of an Existing Study Area

If "New Study Area" is selected, the dialog box as shown in Figure 33 will open for user to specify the file location for the data from the Traffic Information CD, the study area, and the spatial units that will be used in subsequent analysis. After specifying the folder for the traffic data on the Traffic Information CD, the user has the options to select the TTMSs and PTMSs that are located on roadway segments in either a rural or urban area, or both, for analysis. The spatial unit may be district, MPO, county, or route. For example, with the choice of district as the

spatial unit, the dialog box in Figure 34 will be displayed, which allows the MSFs for the TTMSs and PTMSs located within multiple FDOT districts to be retrieved from the Traffic Information CD. Figures 35 and 36 show the dialog boxes from which the user may select MPOs or counties to define the study area.

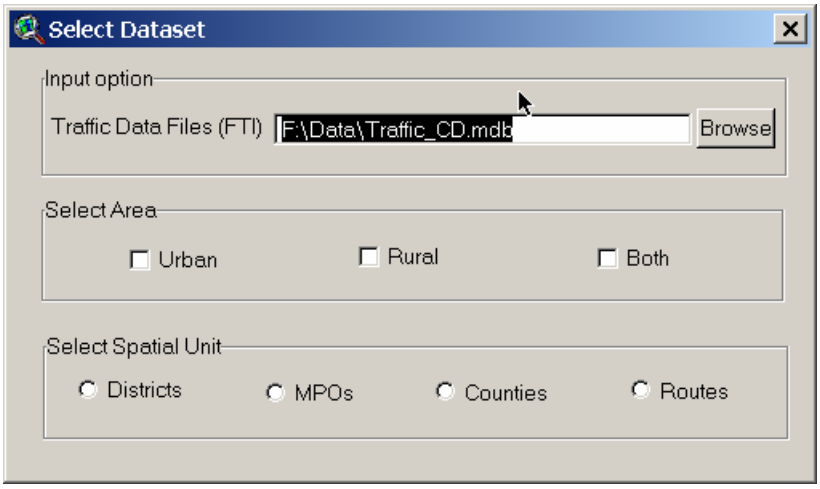


Figure 33. Selecting Dataset Dialog Box

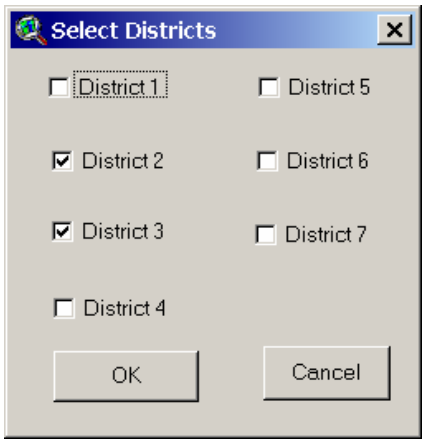


Figure 34. District Selection Dialog Box

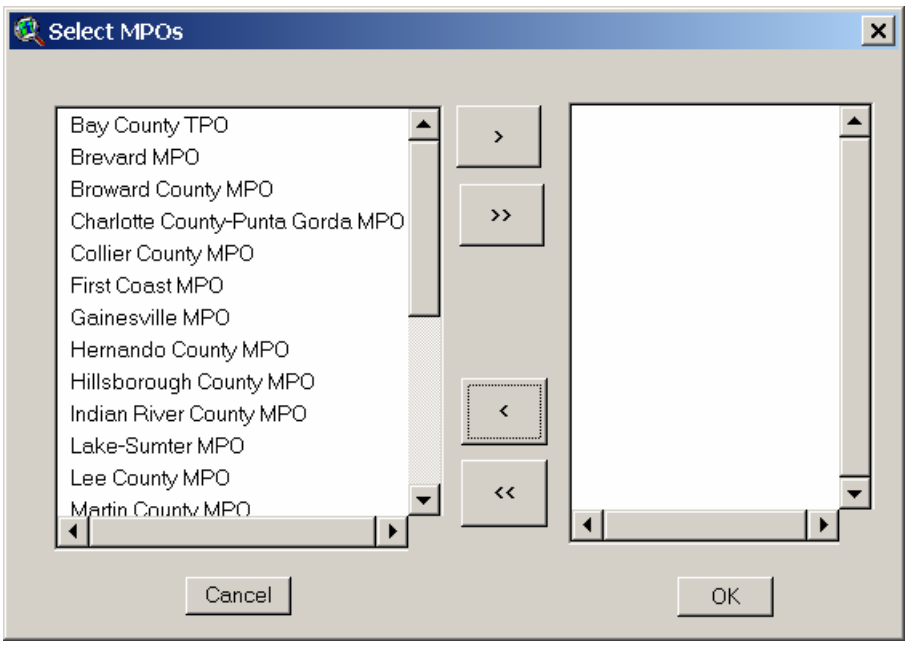


Figure 35. MPO Selection Dialog Box

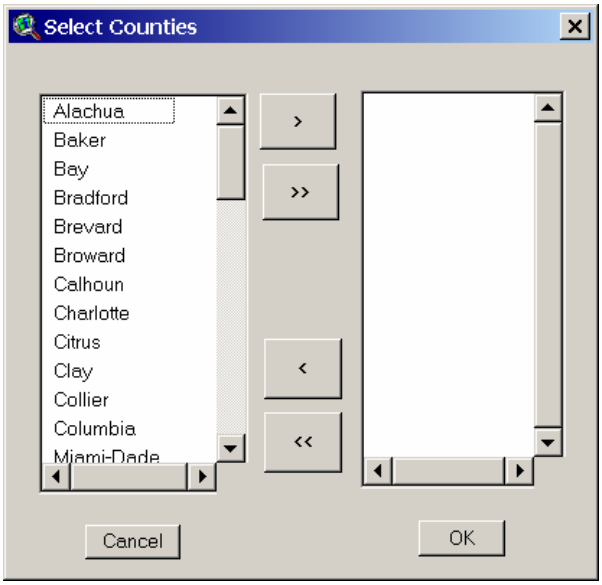


Figure 36. County Selection Dialog Box

Figure 37 shows the five road types that the user may select if the spatial unit is specified as route. They are Interstate, U.S. Highway, State Road, County Road, and Toll Road. The user may choose multiple routes of different types. After the necessary information is provided, the program creates two theme files for the TTMSs and PTMSs located in the study area, which are represented as points on the map with their MSFs as the attributes. Figure 38 shows the TTMSs and PTMSs retrieved from the Traffic Information CD for the tri-county area in southeast Florida.

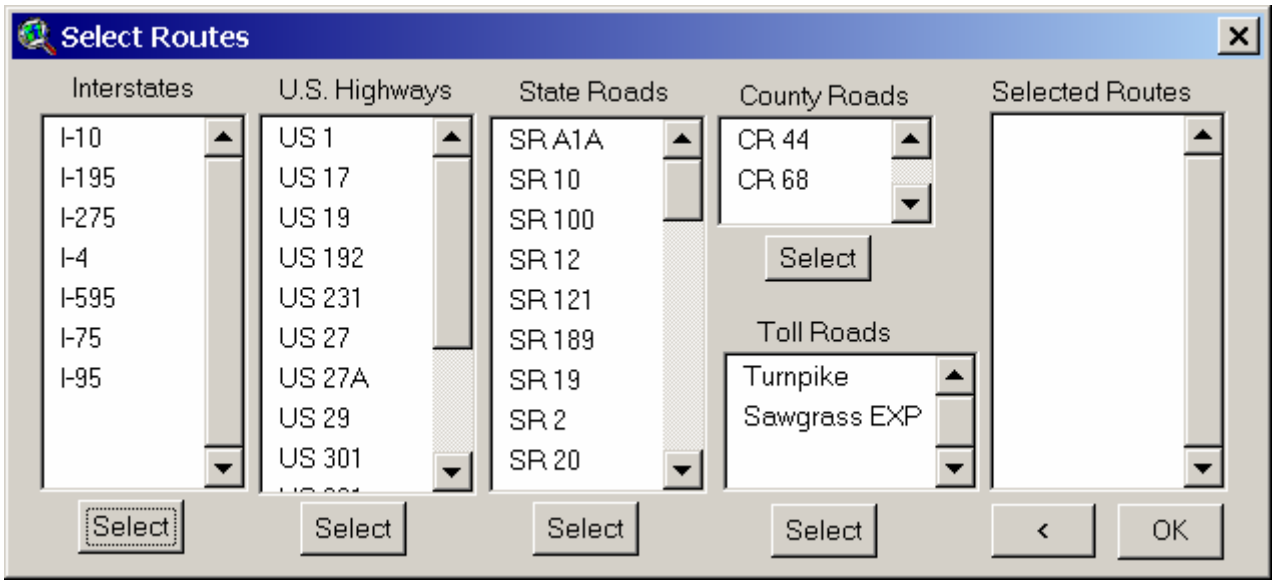


Figure 37. *Select Routes* Dialog Box

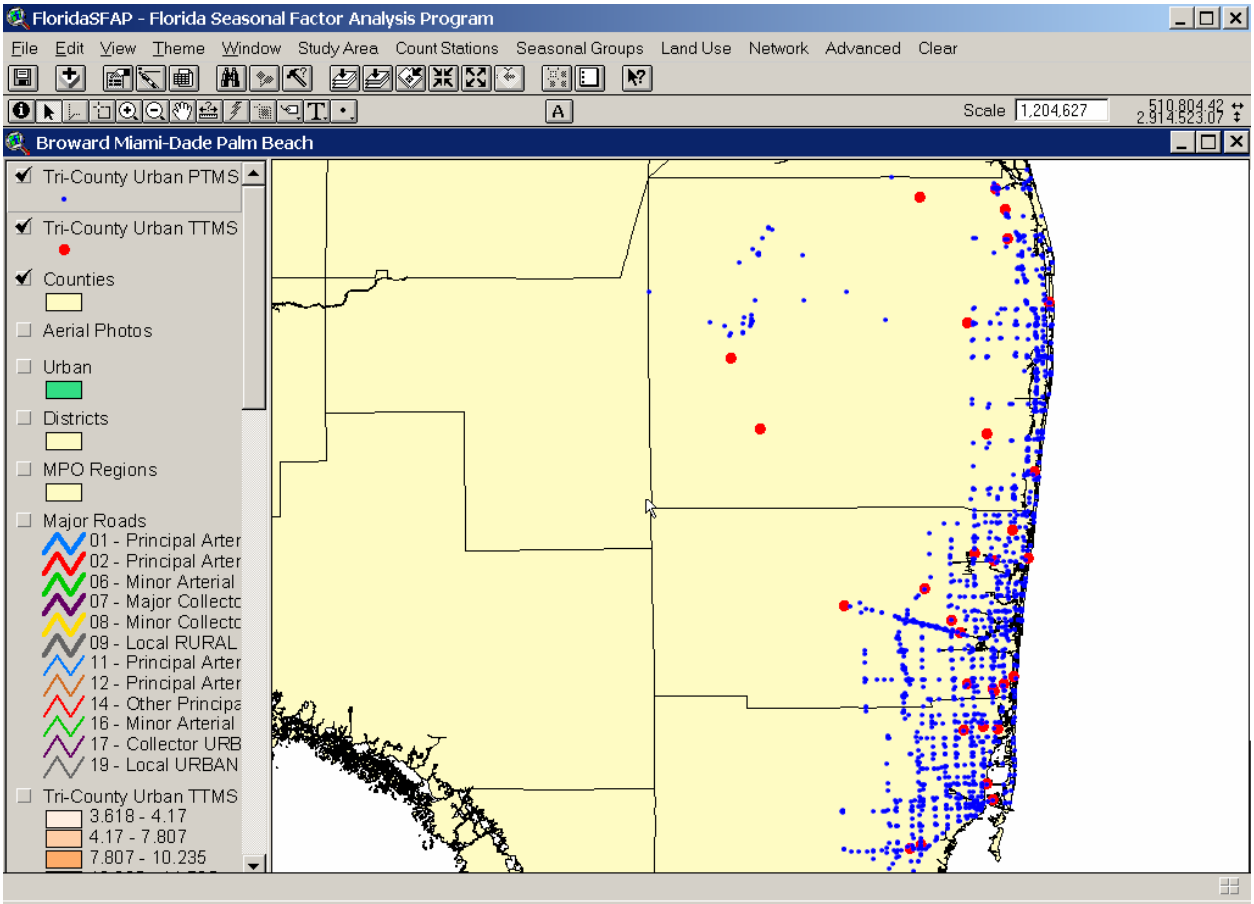


Figure 38. TTMSs and PTMSs in the Tri-County Area

6.2 Count Station Menu

The *Count Station* menu, as illustrated in Figure 39, displays the information in the buffer area of TTMSs, PTMSs, or any location selected by the user. The user may select a specific traffic count station for display by either clicking the count site on the map or typing in the COSITE number, which is the count station identifier (see Figure 40). The user may also point and click on any point on a roadway to display the information in the buffer area for that location (see Figure 41). For this operation, the tool button labeled as "A" must remain pressed. If the user chooses to first zoom in on a roadway, the "A" tool button must be pressed (see Figure 42) to resume the selection process to display the information for other locations.

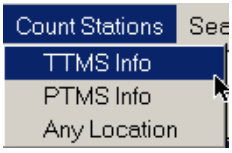


Figure 39. *Count Station* Menu

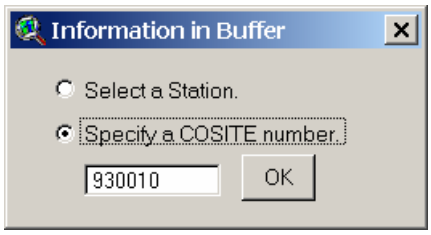


Figure 40. Dialog Box to Select a Count Station for Display

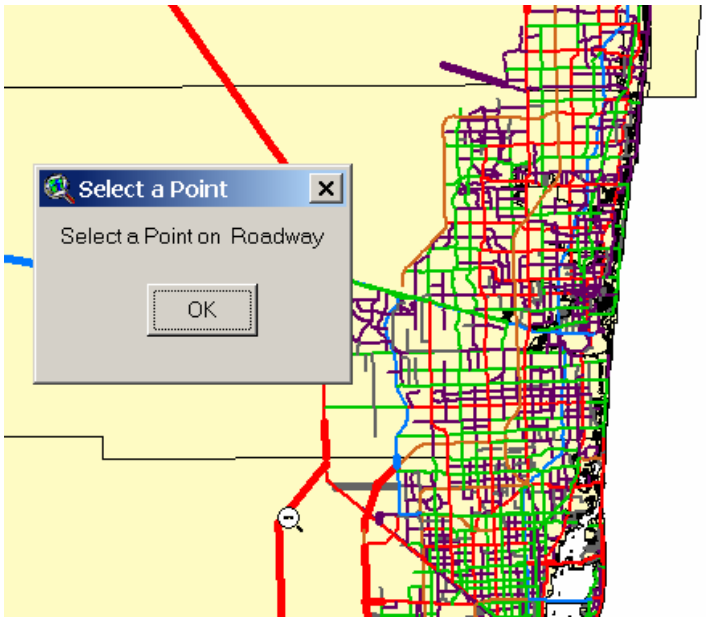


Figure 41. Select Any Location

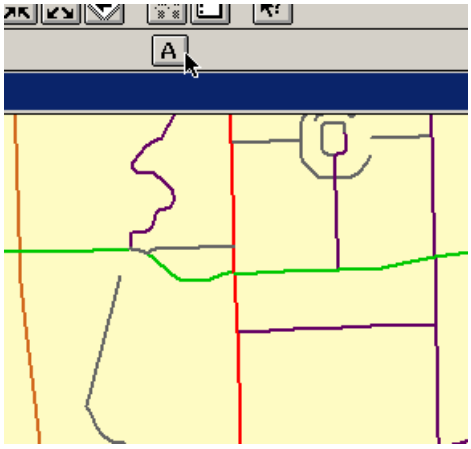


Figure 42. Function Button to Continue Selecting Other Locations for Display

Figures 43 to 45 illustrate the information that is displayed for a selected TTMS, PTMS, and any location on a roadway, respectively. The information includes group number, functional classification, buffer size, AADT, number of lanes, percentage of seasonal households, hotel population, retail employment ratio, and percentage of high-income retired households. For PTMSs and any location on a roadway, the program will also display the information for the closest three TTMSs in the buffer area.

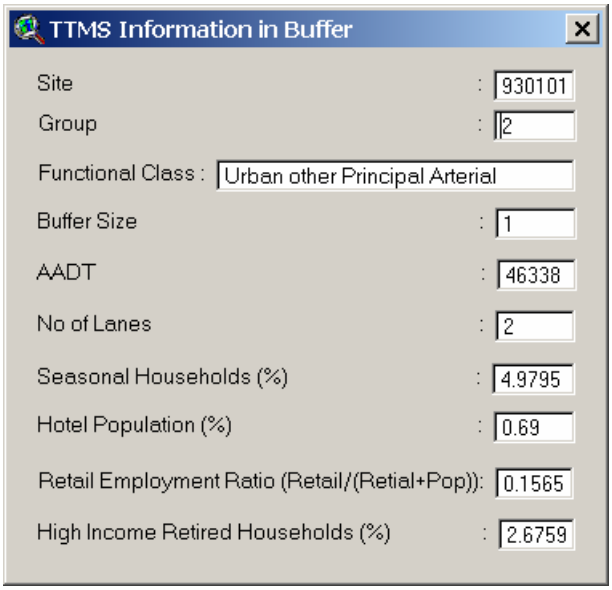


Figure 43. Buffer Information for a Selected TTMS

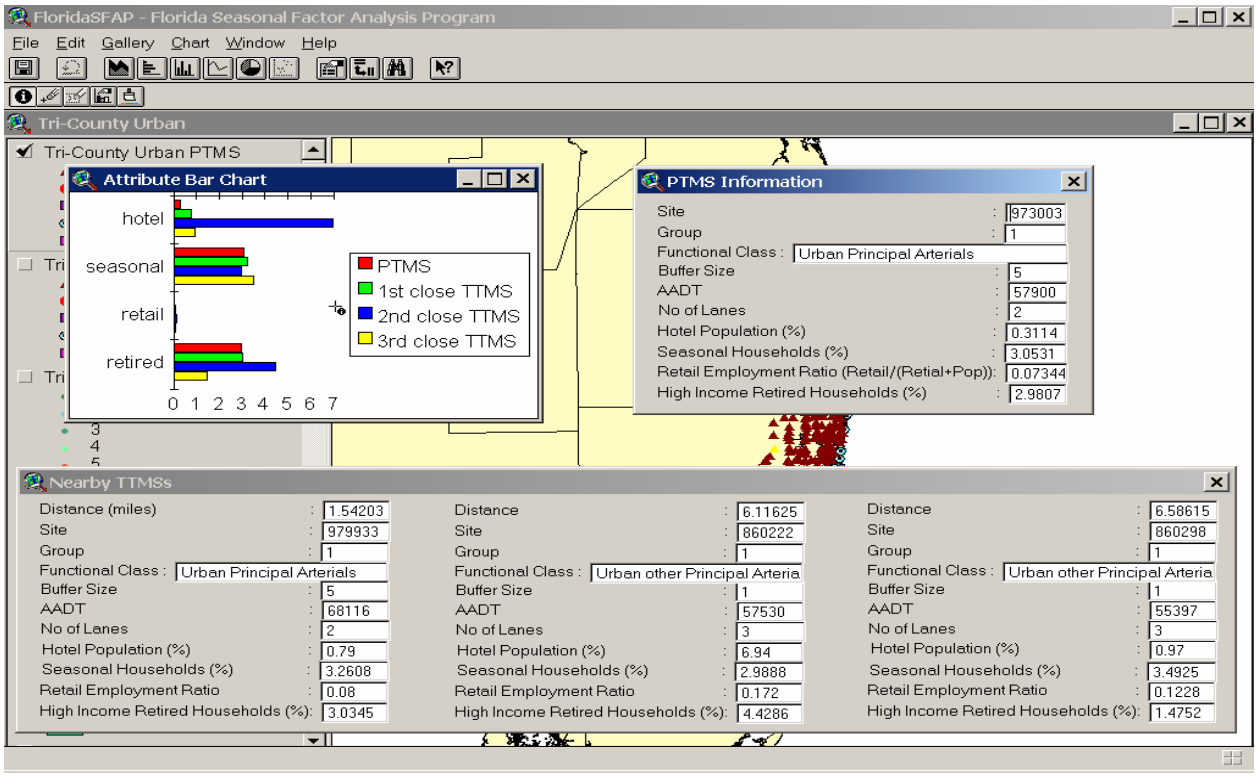


Figure 44. Buffer Information for a Selected PTMS and Three Adjacent TTMSs

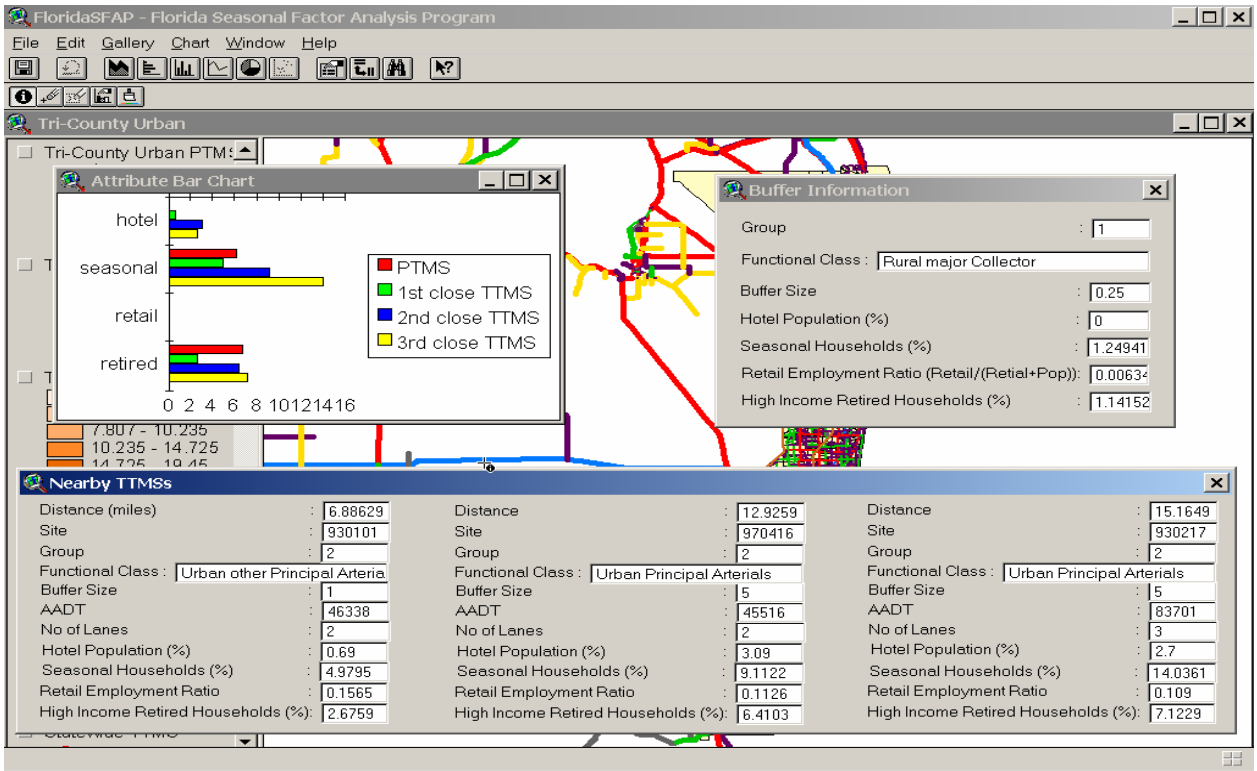


Figure 45. Buffer Information for a Selected Location and Three Adjacent TTMSs

6.3 Seasonal Groups Menu

The *Seasonal Groups* menu, as shown in Figure 46, allows the user to perform the following tasks relevant to the seasonal factor groups: modifying the grouping by removing a TTMS from one group, adding it to another, or creating new groups; graphing the MSF profile for a specific TTMS or for multiple TTMSs that are either selected by the user or are in a given group. The menu allows the information for specific variables, e.g., industrial employment, to be displayed for all buffers of the TTMSs in the study area. Display of contours for seasonal groups may also be requested.

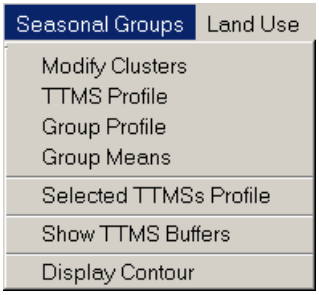


Figure 46. *Seasonal Group* Menu

Figure 47 shows the dialog box for modifying seasonal groups. The user may click the *Show Clusters* button to display the current seasonal factor groups, which may be produced from model-based cluster analysis, hierarchical cluster analysis, or other methods. To modify the groups, click *Select Stations* button to select one or more TTMSs (hold down the Shift key to select multiple TTMSs) and then the "Modify" button to enter the new group number (s).

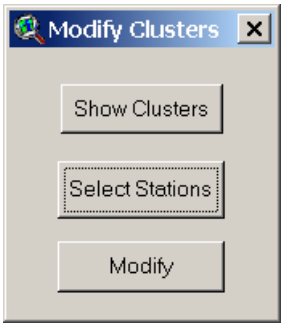


Figure 47. Dialog Box for Cluster Modification

The *TTMS Profile* function on the *Seasonal Groups* menu is designed to display the annual MSF profile, as shown in Figure 48, for a selected TTMS. The *Group Profile* function will display the group statistics for a selected group, which include the group means of the MSFs, the user specified thresholds of the group average, and the minimum and maximum ranges of individual TTMS profiles. Figure 49 illustrates a profile for a given group with ±10% thresholds.

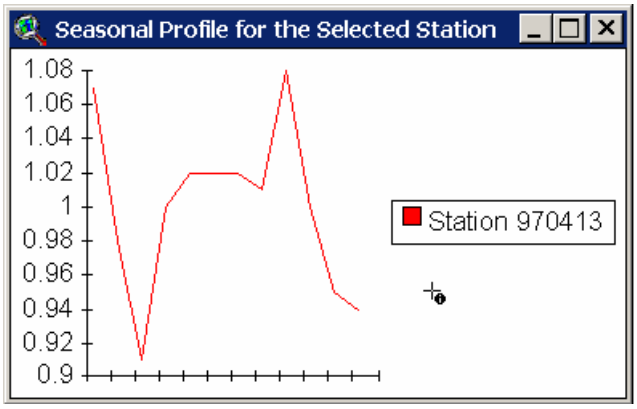


Figure 48. Seasonal Profile for a Selected TTMS

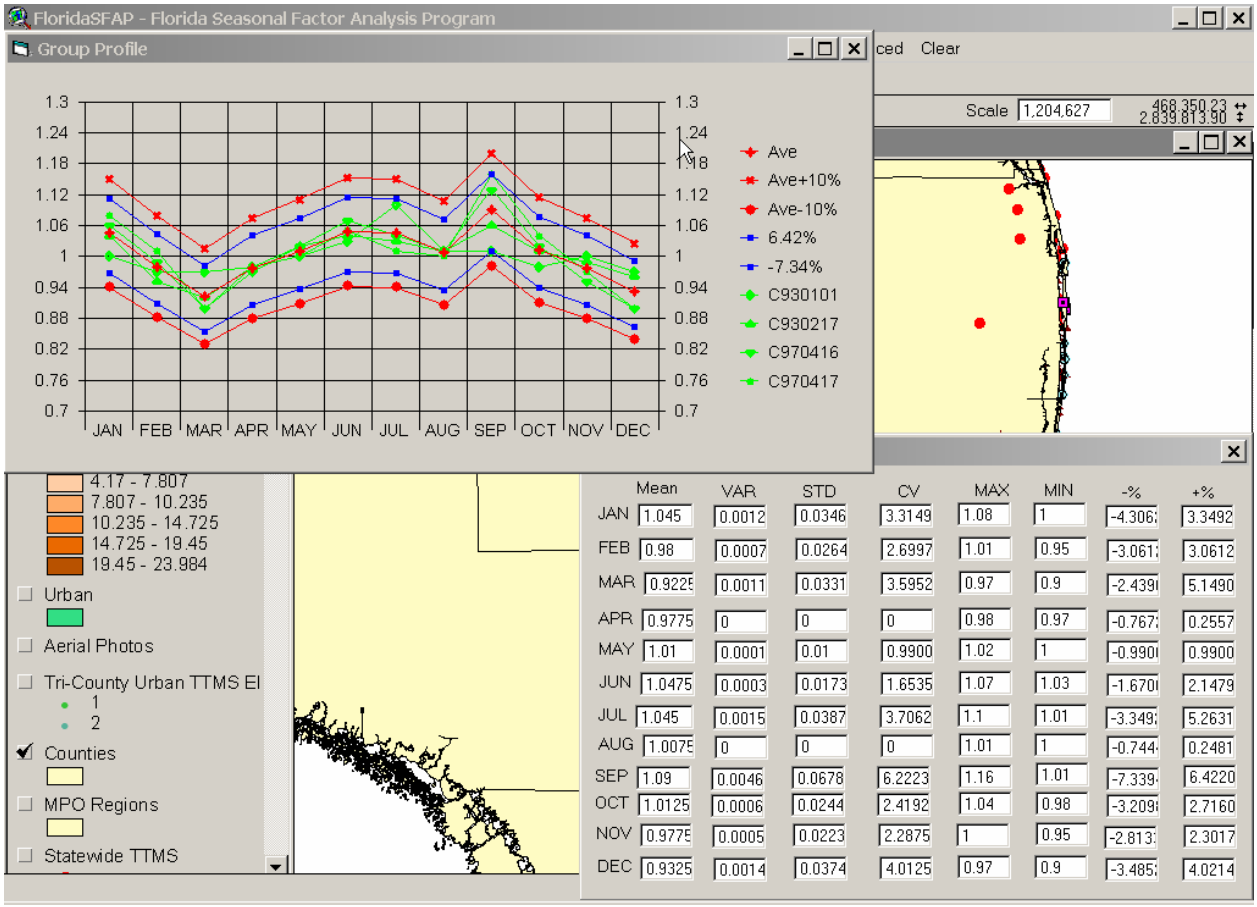


Figure 49. Group Profile and Statistics

Figure 50 shows the MSF group means for all seasonal factor categories in the study area when the "*Group Means*" function as shown in Figure 46 is executed. Figure 51 shows the dialog box if the "*Selected TTMSs Profiles*" function is selected. From this dialog box, the user may select TTMS stations in the study area, followed by a click on the "*OK*" button to visualize these stations' seasonal traffic patterns. Clicking the *Show TTMS Buffers* button will display the

information for a specific variable (see Figure 52 for available variables) for all buffers in the study area (see Figure 53). Clicking the *Display Contour* button will display the contours for the seasonal groups in the study area (see Figure 54).

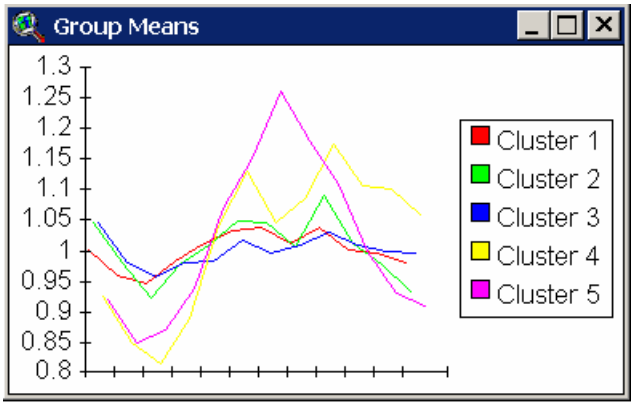


Figure 50. Group Means

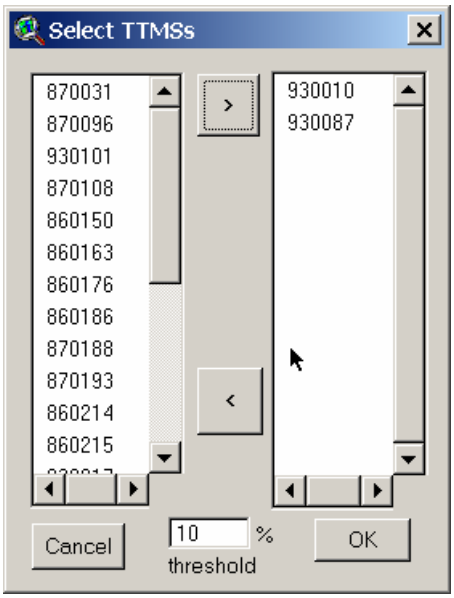


Figure 51. TTMS Selection Dialog Box for MSF Profile

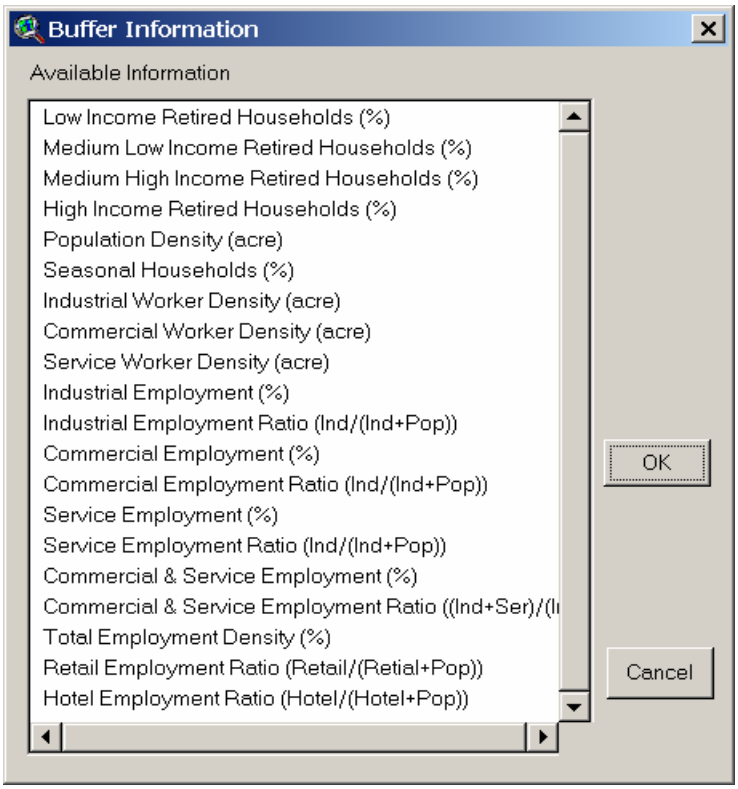


Figure 52. Available Variables for Buffer Information in All Buffers

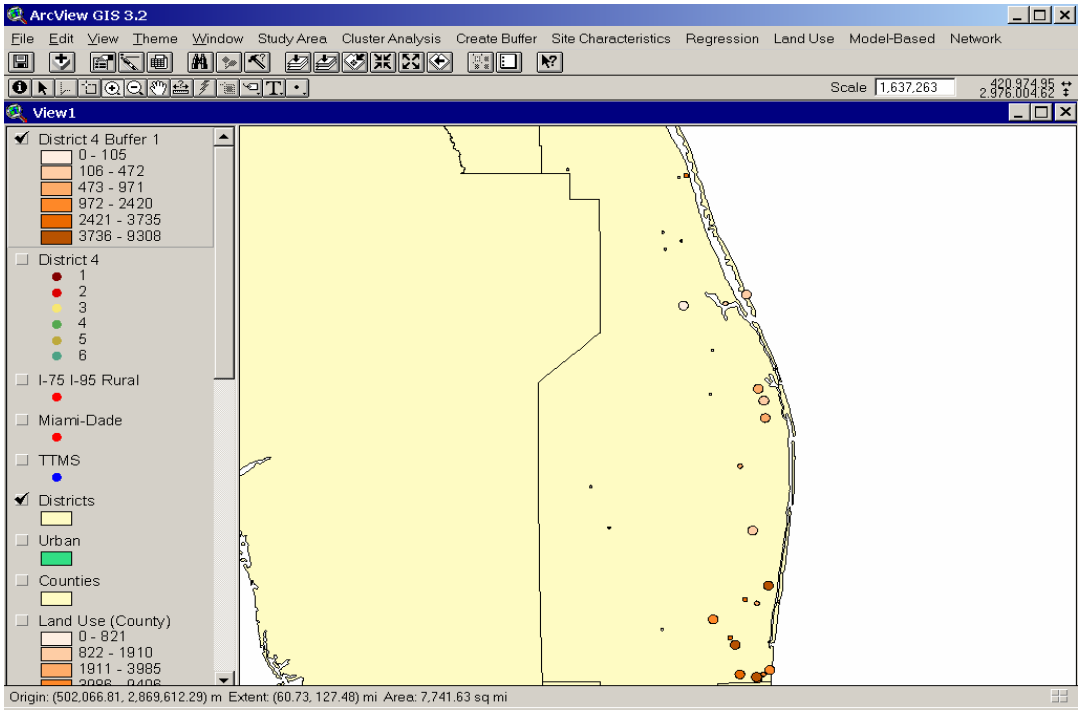


Figure 53. Information for All Buffers

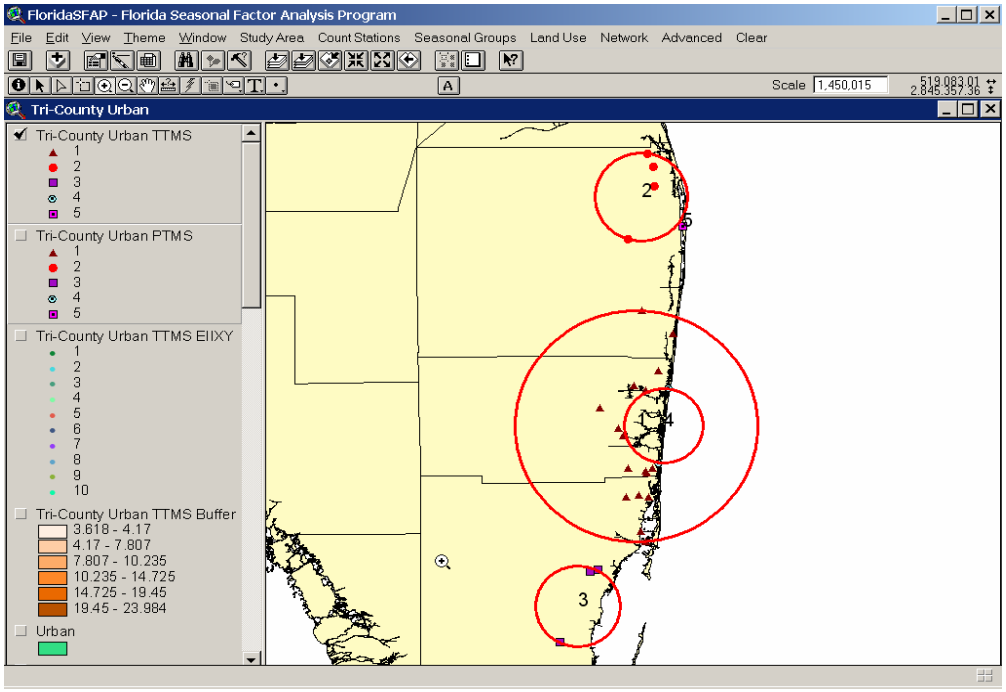


Figure 54. Contours for Seasonal Groups

6.4 Land Use Menu

The *Land Use* menu, as shown in Figure 55, provides options of displaying several typically available land use variables on a GIS map. Currently, 23 land use variables may be displayed. They are:

Aerial Photos. The program displays the one-meter resolution digital aerial photos. Currently, however, only the aerial photos for the Broward County are included.

Set Spatial Unit. As shown in Figure 56, the user may select different spatial units to display land use data. The available units are MPO, county, census tract, census block groups, and TAZ.

Population, Density, Seasonal Household, Median Income, and Retired Population. These five menu entries display the distribution of population, population density, seasonal households, median household income, and retired population data for the selected spatial unit. Figure 57 shows a map displaying the population density per acre by TAZ.

Employment. The employment data retrieved and geocoded from InfoUSA are classified into 15 categories according to the SIC code of a business establishment. The description for each category and the responding SIC code is given in Table 42.

Figure 55. *Land Use* Menu

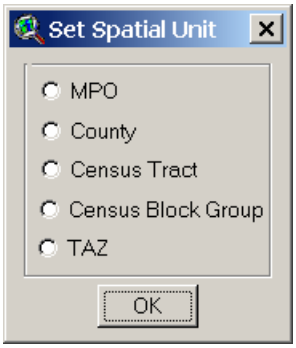


Figure 56. *Select Spatial Unit* for Display Land Use Data

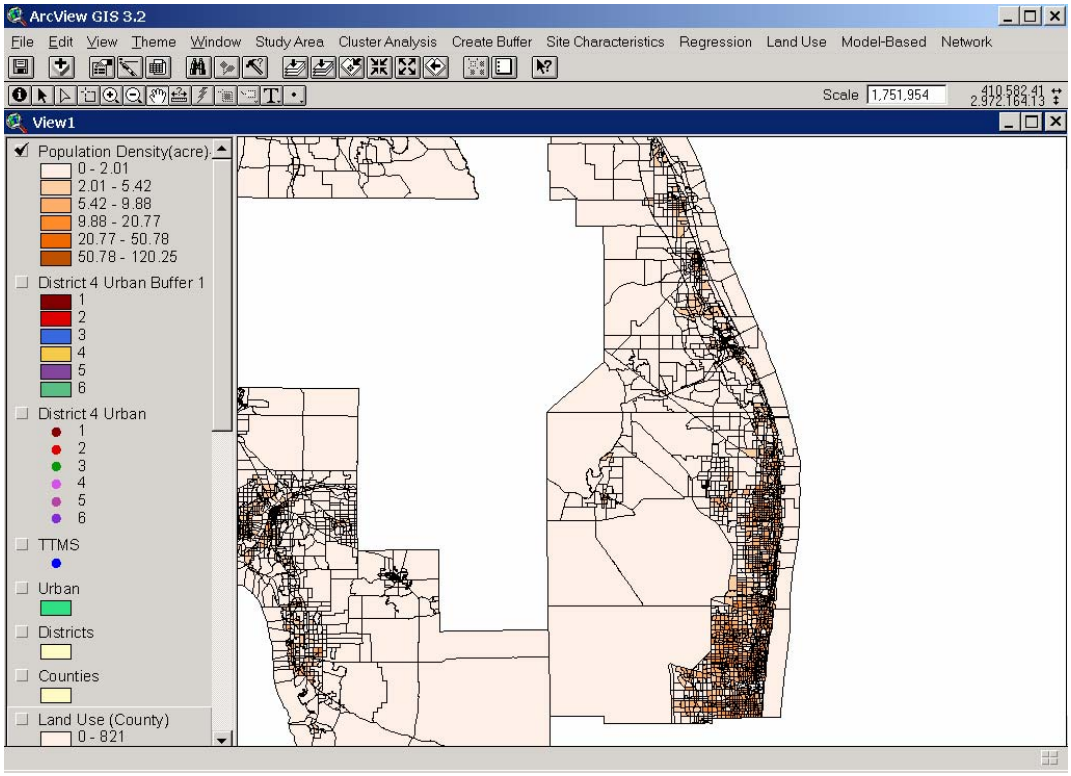


Figure 57. Displaying Population Density for FDOT District 4 by TAZ

Table 42. Description for Employment Category

SIC	Description	Category
1	Agricultural Production-Crops	Agricultural Production-Crops
2	Agricultural Production-Livestock	Agricultural Production-Livestock
7	Agricultural Services	Agricultural Services
8	Forestry	Agricultural Services
9	Fishing Hunting & Trapping	Agricultural Services
10	Metal Mining	Mining
12	Coal Mining	Mining
13	Oil & Gas Extraction	Mining
14	Mining & Quarrying-Nonmetallic Miner	Mining
15	Building Construction-General Contractor	Construction
16	Building Construction-General Contractor	Construction
17	Construction-Special Trade Contractor	Construction
20	Food & Kindred Products Manufacture	Manufacturing
21	Tobacco Products Manufacturing	Manufacturing
22	Textile Mill Products Manufacturing	Manufacturing
23	Apparel & Other Finished Products Manufacturing	Manufacturing
24	Lumber & Wood Prods Except Furniture Manufacturing	Manufacturing
25	Furniture & Fixtures Manufacturing	Manufacturing
26	Paper & Allied Products Manufacturing	Manufacturing
27	Printing Publishing & Allied Industry	Manufacturing
28	Chemicals & Allied Products Manufacturing	Manufacturing
29	Petroleum Refining & Related Industry Manufacturing	Manufacturing
30	Rubber & Miscellaneous Plastics Manufacturing	Manufacturing
31	Leather & Leather Products Manufacturing	Manufacturing
32	Stone Clay Glass & Concrete Products Manufacturing	Manufacturing
33	Primary Metal Industries Manufacturing	Manufacturing
34	Fabricated Metal Products Manufacturing	Manufacturing
35	Industrial & Commercial Machinery Manufacturing	Manufacturing
36	Electronic & Other Electrical Equipment	Manufacturing
37	Transportation Equipment Manufacturing	Manufacturing
38	Measuring & Analyzing Instruments Manufacturing	Manufacturing
39	Miscellaneous Industries Manufacturing	Manufacturing
40	Railroad Transportation	Transportation
41	Local/Suburban Transit & Highway Passenger	Transportation
42	Motor Freight Transportation/Warehouse	Transportation

SIC	Description	Category
43	United States Postal Service	Transportation
44	Water Transportation	Transportation
45	Transportation by Air	Transportation
46	Pipelines Except Natural Gas	Transportation
47	Transportation Services	Transportation
48	Communications	Transportation
49	Electric Gas & Sanitary Services	Transportation
50	Wholesale Trade-Durable Goods	Wholesale
51	Wholesale Trade-Nondurable Goods	Wholesale
52	Building Materials & Hardware	Retail
53	General Merchandise Stores	Retail
54	Food Stores	Retail
55	Automotive Dealers & Service Station	Retail
56	Apparel & Accessory Stores	Retail
57	Home Furniture & Furnishings Stores	Retail
58	Eating & Drinking Places	Retail
59	Miscellaneous Retail	Retail
60	Depository Institutions	Insurance & Real Estate
61	Non-depository Credit Institutions	Insurance & Real Estate
62	Security & Commodity Brokers	Insurance & Real Estate
63	Insurance Carriers	Insurance & Real Estate
64	Insurance Agents Brokers & Service	Insurance & Real Estate
65	Real Estate	Insurance & Real Estate
67	Holding & Other Investment Offices	Insurance & Real Estate
70	Hotels Rooming Houses & Camps	Hotels & Camps
72	Personal Services	General Services
73	Business Services	General Services
75	Auto Repair Services & Parking	General Services
78	Motion Pictures	General Services
80	Health Services	General Services
81	Legal Services	General Services
82	Educational Services	General Services
83	Social Services	General Services
87	Engineering & Accounting & management Services	General Services
88	Private Households	General Services
89	Miscellaneous Services	General Services
79	Amusement & Recreation Services	Recreation Services
84	Museums Art Galleries & Gardens	Recreation Services
86	Membership Organizations	Recreation Services
91	Executive Legislative & General Government	Public Administration
92	Justice Public Order & Safety	Public Administration

SIC	Description	Category
93	Public Finance & Taxation Policy	Public Administration
94	Administration-Human Resource Programs	Public Administration
95	Admin-Environmental Quality Programs	Public Administration
96	Administration of Economic Programs	Public Administration
97	National Security & International Affair	National Security

6.5 Network Menu

As illustrated in Figure 58, the *Network* menu allows the user to display the functional classifications of the roadways in the study area. An example of a street network with functional classification information is given in Figure 59.

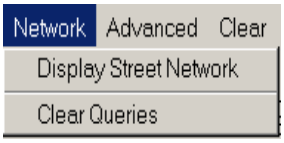


Figure 58. *Network* Menu

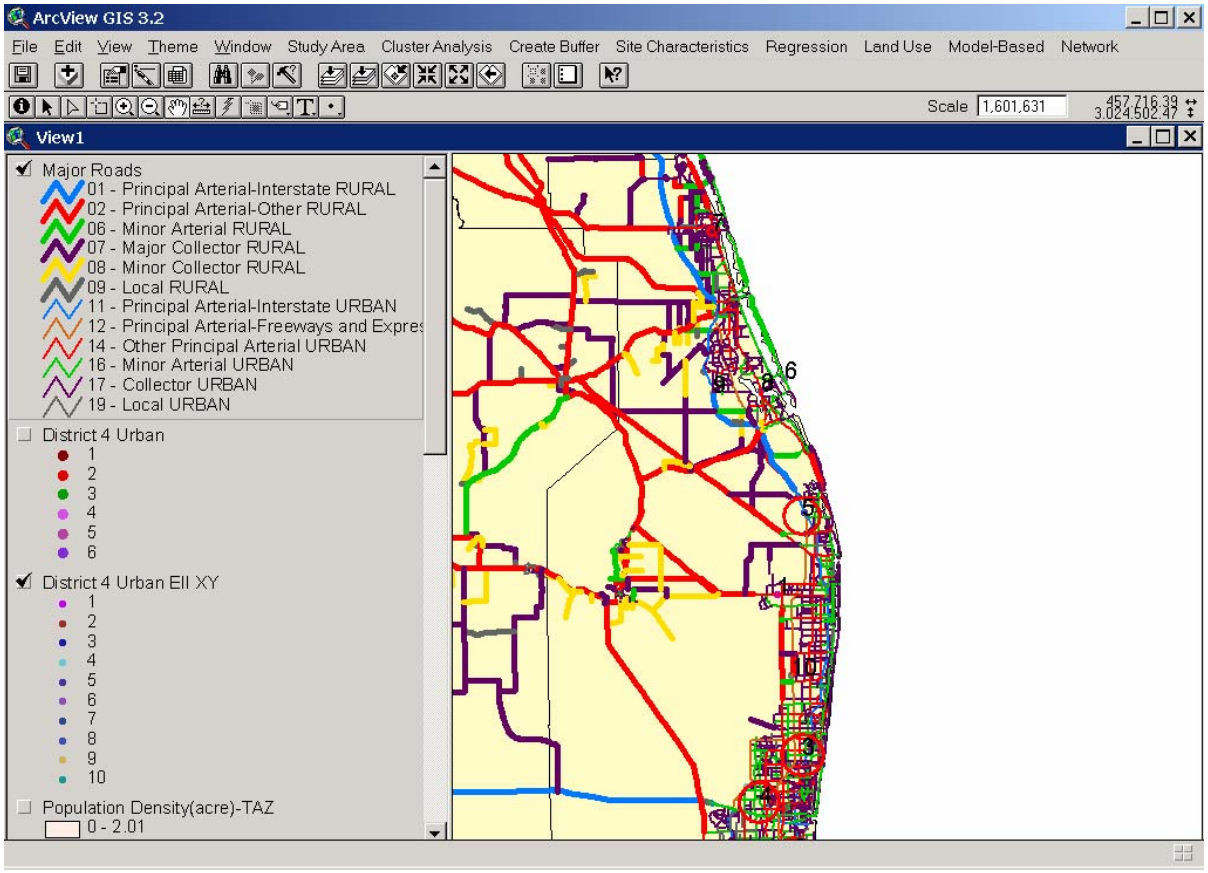


Figure 59. Functional Classifications for Roadways

6.6 Advanced Menu

The *Advanced* menu, as shown in Figure 60, allows the user to perform several advanced analyses, including cluster analysis, buffer analysis, regression analysis, and model-base cluster analysis. The *Cluster Analysis* function allows the user to perform hierarchical cluster analysis using different methods available in the SAS program. Currently, eight clustering methods are available in the program and the user may specify which model to use in the *Select Method* dialog box, where Average Linkage is the default (see Figure 61). To run cluster analysis in SAS, the user needs to specify the number of clusters desired for hierarchical cluster analysis. An example output for six clusters is shown in Figure 62.

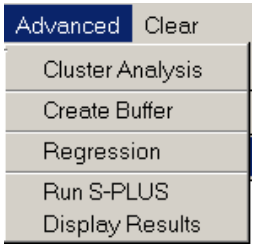


Figure 60. *Advanced* Menu

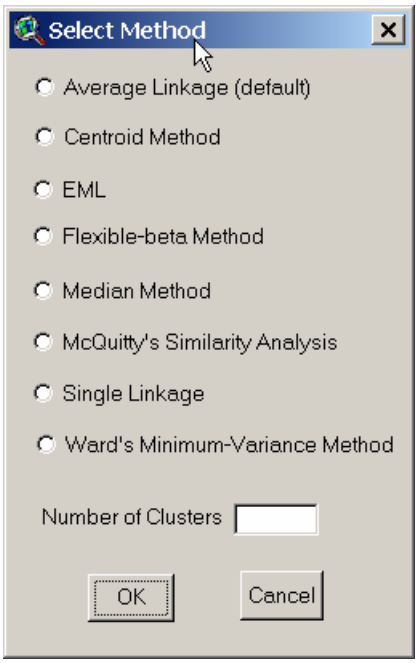


Figure 61. Clustering Methods

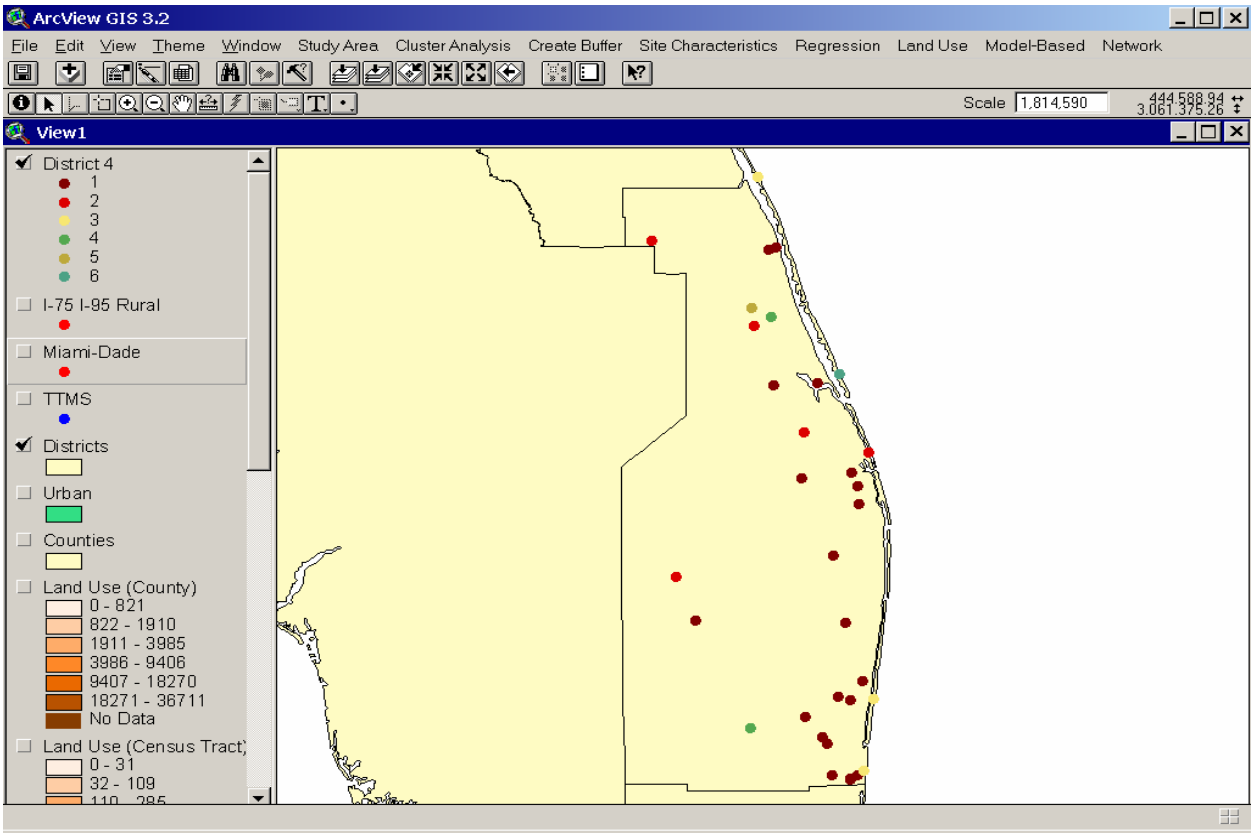


Figure 62. Cluster Results from SAS

The *Create Buffer* function allows the user to compile data for the buffer area of a TTMS. The data may be visualized and incorporated into regression analysis. Figure 63 shows the dialog box for the user to specify the buffer size for different roadway types. The values displayed are the default values. These values also universally apply to the entire study area.

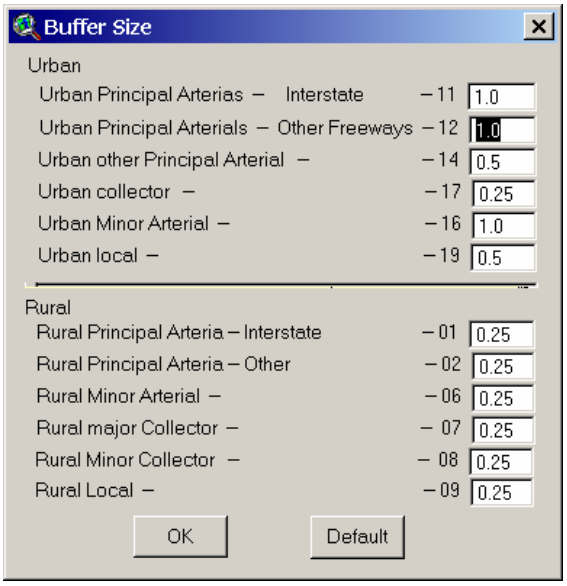


Figure 63. Dialog Box for Specifying Buffer Size

The user may also choose data at either the TAZ or Census block group level, as shown in Figure 64. The result of this function will be a map displaying the buffers created for the study area (see Figure 65).

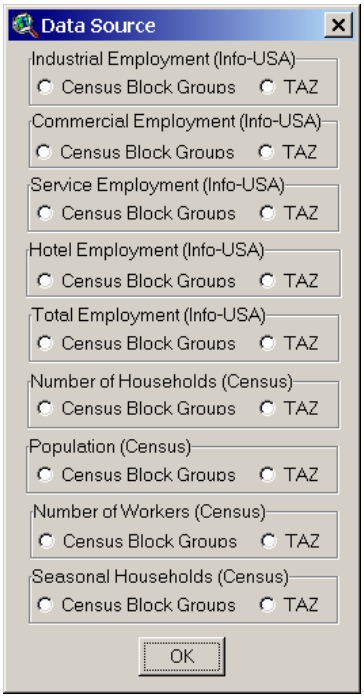


Figure 64. Data Source for Data Compilation

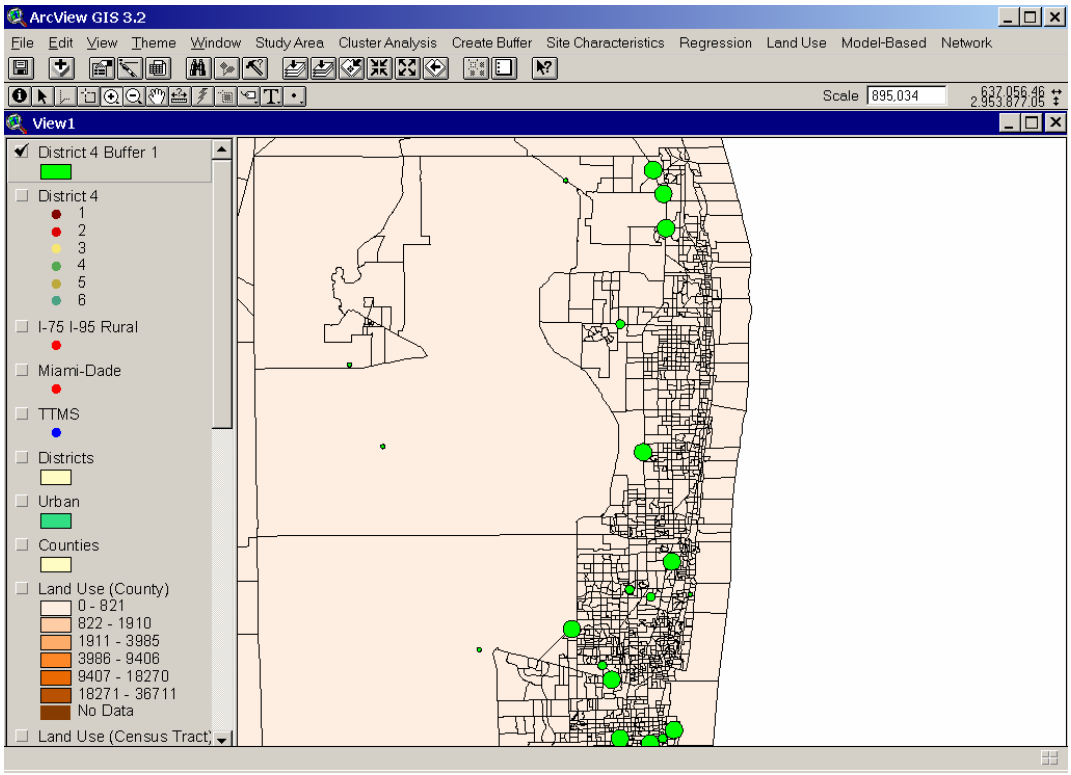


Figure 65. Example for Creating Buffer

The program also allows the user to perform regression analysis as described in Sections 5.1 and 5.2 with the *Regression* function. The user may select multiple variables from the list of available regressors (see Figure 66) and view the output (see Figure 67).

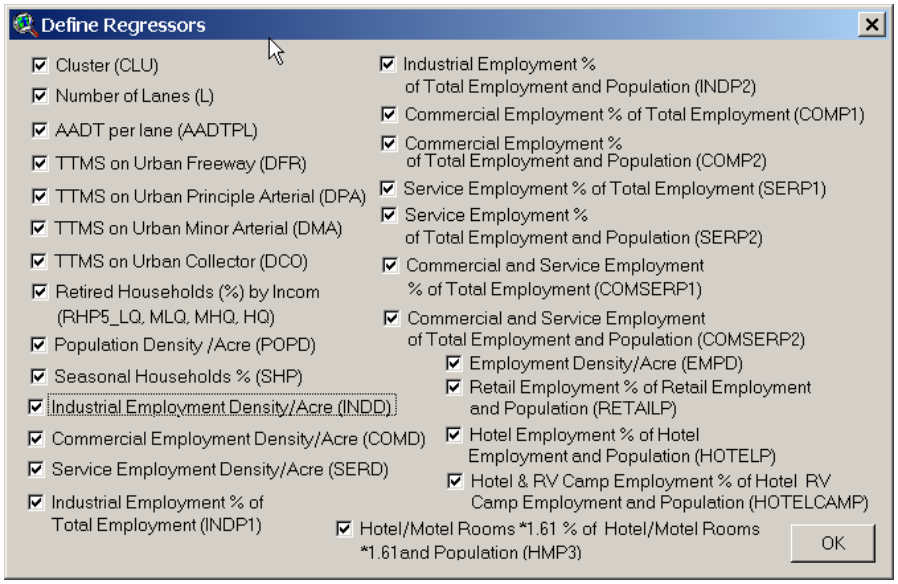


Figure 66. Selection of Regressors for Regression Analysis

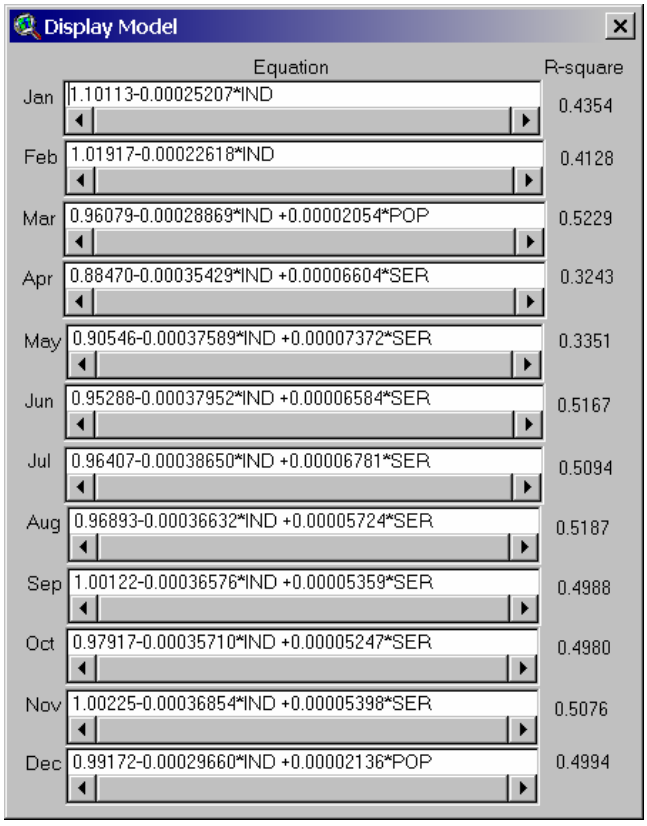


Figure 67. Regression Analysis Results from SAS

The *Run S-PLUS* menu allows the user to select one of the 10 parametric cluster methods available in MCLUST as previously described in Section 2.2.3. Figure 68 shows the *Model-Based* menu. The check box provides the user with the option to include the geographical coordinates of the TTMSs to be simultaneously considered with the MSF patterns during the grouping process. The user is expected to choose a method such as the EEI method, which is recommended. Clicking the *OK* button will automatically trigger S-Plus, attach MCLUST, and then run the model specified. Selecting the *Display Results* entry (see Figure 60) will display the dialog box shown in Figure 69 to let the user choose the grouping results to be displayed after the S-Plus run is completed and the grouping results are saved in a dBase file. As shown in figure 69, the dialog box will show the optimal number of groups for a given run. The user may select the number of groups to view. The *Display Contour* option will display the grouping results with circles drawn to show the area occupied by each seasonal factor group. The radius is determined by the distance between the centroid of the TTMS in the group and the TTMS in the group that is farthest from the centroid. Figure 70 shows an example of contours for the model-based cluster analysis. The check box for *Display Profile* will allow the MSF profile and the 10% thresholds of the group average to be displayed.

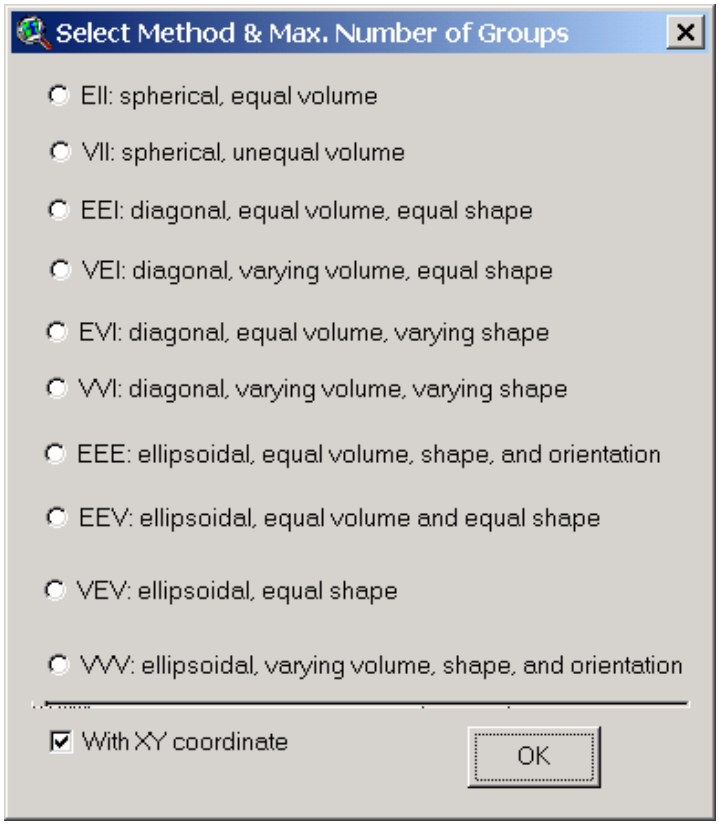


Figure 68. Models Available in MCLUST

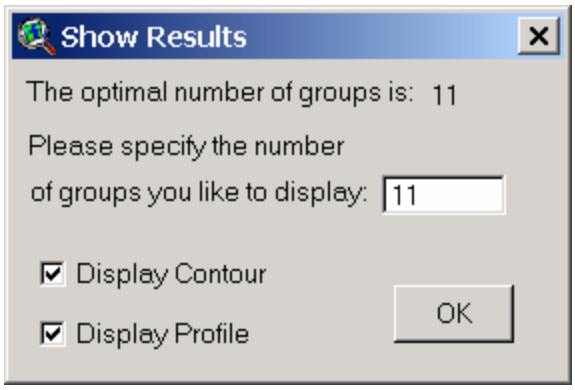


Figure 69. Show Results Dialog Box

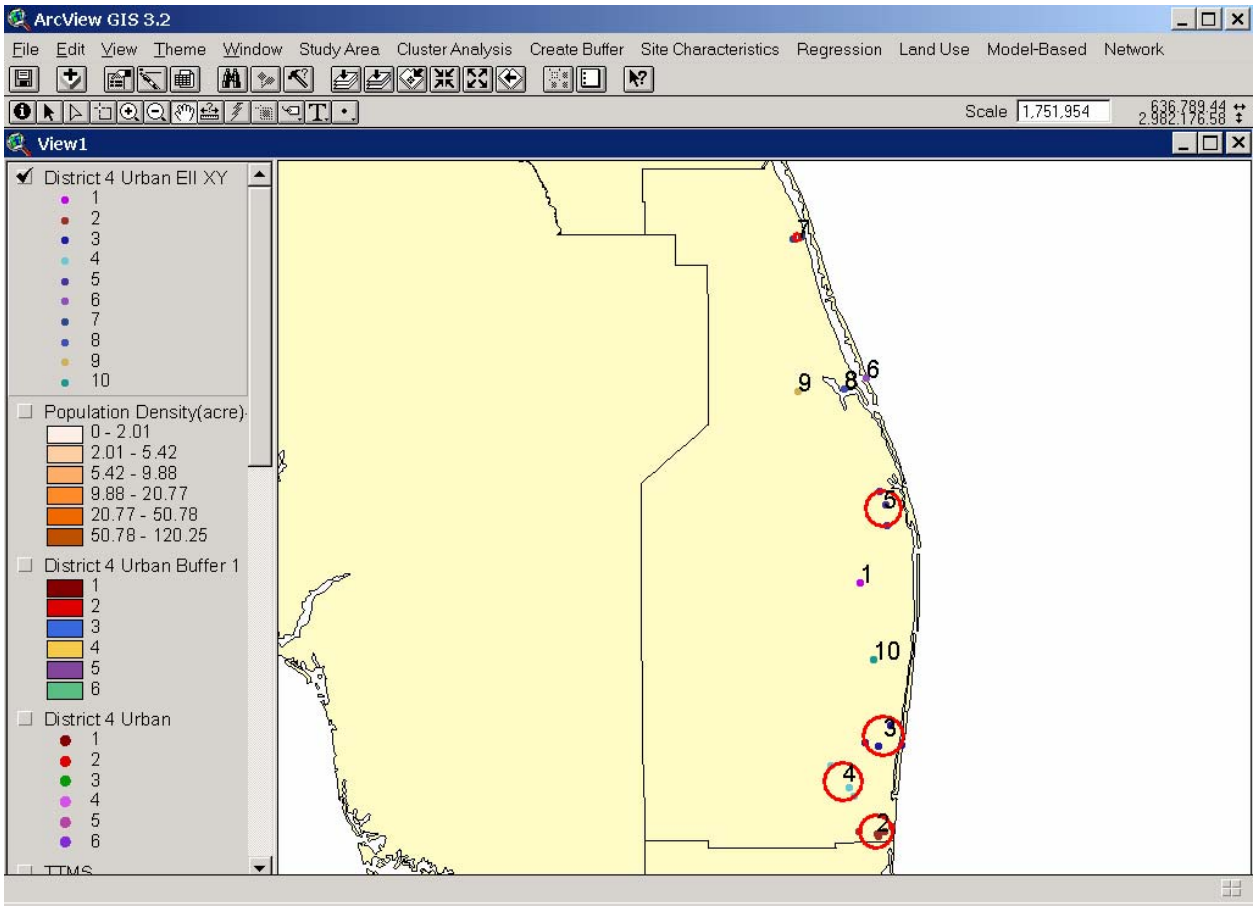


Figure 70. Group Contours

7. CONCLUSIONS AND RECOMMENDATIONS

Seasonal factors are a complex subject. While there have been relatively more studies on various methods to determine seasonal groups, determining the underlying causes of season variations in traffic and developing models to predict seasonal groups has proven to be a significant challenge. So far based on literature, success in explaining or modeling seasonal factors has been limited. This research made contributions to the understanding of the subject by identifying plausible predictors for seasonal groups, further confirming the importance of geographic location in seasonal grouping, and providing a theoretical basis for consideration of geographic locations, and developing a practical approach for assigning short counts to seasonal groups.

This study first investigated conventional nonparametric hierarchical clustering analysis and parametric model-based cluster analysis methods for seasonal factor grouping. The results from the hierarchical clustering analyses suggested that spatial proximity should be appropriately considered in both grouping and assignment processes. The model-based clustering analyses provided a good starting point for transportation professionals to more accurately group TTMSs into seasonal factor categories in a systematic and data-driven manner by simultaneously considering a TTMS's spatial proximity and their MSFs. This grouping approach was incorporated in this study to reduce the time and effort in grouping TTMSs for seasonal factor categories.

Multiple linear regression analyses were subsequently conducted for selected urban and rural areas to identify possible explanatory variables for seasonal traffic fluctuations. The regression models considered the effect of spatial proximity by introducing the locations of the count stations into the models. The regression models calibrated with the MSFs collected from the TTMSs located on the urban roads in southeast Florida explained significantly more variations in the data than the rural models did. One possible explanation for such a discrepancy between the urban and rural models is that more detailed and accurate land use, socioeconomic, and demographic data were available for urban areas. Additionally, the buffer method employed in this study is unable to describe adequately the land use, socioeconomic, and demographic characteristics of through traffic, which is not originated or destined in local buffer areas. The higher the function class of a road is, the more significant through traffic will be, especially on rural roads. Seasonal residents, tourists, retired people between age 65 and 75 with high income, and retail employment were identified as the significant indicators for seasonal traffic fluctuation on urban roads in southeast Florida. For the rural roads, variables such as functional classification for highways, percentage of seasonal households, agricultural employment, and truck factor were identified as potential explanatory variables.

To develop a methodology to assign a seasonal factor category to a PTMS, a fuzzy decision tree was constructed using the TTMS groups obtained from the model-based cluster analysis and based on the aforementioned four variables, i.e., ratio of seasonal households to permanent households (*SHP*), hotel population to hotel population plus households population (*HMP3*), retail workers as a percentage of total retail workers plus population (*RETAILP*), and percentage of retired households of the highest income quartile (*RH5_HQ*) for the tri-county urban area, i.e., Broward, Miami-Dade, and Palm Beach counties. The decision tree was then applied to determine the seasonal factor category for a given PTMS. The fundamental assumption for such

an application is that the socioeconomic/demographic variables would have the same or very similar effect on traffic fluctuations at the TTMS and PTMS locations on urban roads. The decision tree is easy to visualize and apply, and the assignment results are self-explanatory. For example, areas with a larger number of visitors and a larger number of seasonal households would expect to experience more fluctuation in traffic volumes. It needs to be pointed out, however, that there was still fuzziness in the assignment results due to the fact that the four land use variables did not completely explain the traffic variations, that the sample size was limited, that TTMSs might not have reflected all representative land use patterns, and the membership of a PTMS in a given seasonal factor group might be less than 1 (partial membership). For these reasons, the assignment methodology and results do not entirely replace the transportation analyst, who should examine the results, check the data, and determine if the assignment is reasonable. Some additional data collection (e.g., monthly short counts) may also be necessary to verify the assigned seasonal factor category if the traffic volume is high and the impact of estimated AADT on transportation projects is significant.

A GIS based computer program was developed as part of this research to demonstrate the usefulness of a GIS user interface for visualization of land use, demographic, and socioeconomic data, as well as the characteristics of the transportation systems and traffic counts. Buffer analysis, regression analysis, and cluster analysis were also supported in the program for advanced users who are interested in performing statistical analysis. The statistical functions were provided by SAS and S-Plus.

Although this study developed regression models that could potentially be used to estimate seasonal factors directly for a PTMS, because of the limited sample size, the predictive power of the models could not be determined. Additionally, because traffic in different urban areas may have different seasonal patterns due to differences in climate, local economy, and demographics, variables identified in this study may not be directly applicable to other areas.

The following recommendations were made based on the findings from this research:

- To make the results from this research useful to all FDOT districts, where the seasonal categories are determined and assigned to PTMSs, and even to local government users who operate a local traffic statistics program, additional studies need to be carried out to determine whether the variables identified in this study for the urban areas in southeast Florida are also applicable to other urban areas in the state. Due to differences in local land use patterns and economies, it is possible that some urban areas have a different set of variables that explain the patterns of traffic variations.

- The regression models for estimating MSFs for rural roads currently have relatively low R^2s. To improve the model performance and identify better MSF predictors, further analyses are necessary. They may include the development and testing of improved or new variables and new modeling techniques such as nonlinear regression models.

- A standard procedure should be developed by FDOT based on the results from this study and future studies. This standard procedure should be based on a set of statistics based methods for seasonal factor grouping and assignment that are more objective and data-

driven and that minimize the reliance on individuals' experience and subjective judgment. Such a standard procedure will help improve the quality of the transportation data used in important decision making processes.

- The current prototype GIS program is a demonstration program developed for FDOT District 4. It needs to be expanded to include all FDOT districts. The program and the necessary data need to be delivered in a single CD-ROM, similar to the current traffic CD-ROM published by FDOT each year. The data required by the program, which are from the U.S. Census Bureau and from urban area travel demand models, need to be made available from the Internet. A possible central depository location may be the Florida Geographic Digital Library (FGDL) at the University of Florida. The data should be updated every three to five years as more recent data become available or when census data are released.

- The current GIS program is implemented in the ArcView environment. When the FDOT district offices and central office completely migrate to ArcGIS, this program may be re-implemented by customizing ArcGIS with VBA (the programming language in ArcGIS). Alternatively, a program implemented in MapObject (also an ESRI product) may be developed. The advantage of a MapObject based program is that it does not require any GIS software from the user and still provides the same GIS functionalities. A MapObject based program will allow the GIS program to be distributed to the users on a CD and used in the same way as the Traffic Data CD.

REFERENCES

[AAS92] *AASHTO Guidelines for Traffic Data Programs*, American Association of State Highway and Transportation Officials, Washington, D.C., 1992

[ALB91] Albright, D., "An Imperative for, and Current Progress toward National Traffic Monitoring Standards," *ITE Journal*, June 1991, pp. 22-26.

[ALB93] Albright, D., "Standards, Innovation, and the Future of Traffic Monitoring," *ITE Journal*, January 1993, pp. 31-36.

[AUN00] Aunet, B., "Wisconsin's Approach to Variation in Traffic Data," *North American Travel Monitoring Exhibition and Conference CD*, Wisconsin Department of Transportation, Madison, Wisconsin, August 2000.

[BAN93] Banfield, J.D. and A.E. Raftery, "Model-Based Gaussian and Non-Gaussian Clustering, *Biometrics*, Vol. 49, No. 3, 1993, pp. 803-821.

[BEL78] Bellamy, P.H. *Seasonal Variations in Traffic Flows*, Supplementary Report 437, prepared for the Department of the Environment and the Department of Transport, Prepared by Traffic Engineering Department, Transport and Road Research Labroratory, Berkshire, Great Britain, 1978.

[BPR65] *Guide for Traffic Volume Counting Manual*, 2nd Edition, Bureau of Public Roads, U.S. Department of Commerce, 1965.

[CAR88] Carpenter, G.A. and S. Grossberg, "The Art of Adaptive Pattern Recognition by a Self-Organizing Neural Network," *Computer*, No. 3, Vol. 21, 1988, pp. 77-88.

[CHU98] Chung, J-H, K. Viswanathan, and K.G. Goulias, "Design of Automatic Comprehensive Traffic Data Management System for Pennsylvania," *Transportation Research Record 1625*, Transportation Research Board, National Research Council, Washington, D.C., 1998, pp. 1-11.

[DAV96] Davis, G.A. and Y. Guan, "Bayesian Assignment of Coverage Count Locations to Factor Groups and Estimation of Mean Daily Traffic," *Transportation Research Record 1542*, Transportation Research Board, National Research Council, Washington, D.C., 1996, pp. 30-37.

[DAV97] Davis, G.A., *Estimation Theory Approach to Monitoring and Updating Average Daily Traffic*, Final Report, Minnesota Department of Transportation, St. Paul, Minnesota, January 1997.

[DUN02] Dundar, M.M. and D. Landgrebe, "A Model Based Mixture Supervised Classification Approach in Hyperspectral Data Analysis," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 40, No. 12, December 2002, pp 2692-2699.

[ERH91] Erhunmwunsee, P.O., "Estimating Average Annual Daily Traffic Flow from Short Period Counts," *ITE Journal*, November 1991, pp. 23-30.

[FAG86] Faghri, A., M. Glaubitz, and J. Parameswaran. Development of Integrated Traffic Monitoring System for Delaware, *Transportation Research Record 1536*, Transportation Research Board, National Research Council, Washington, D.C., 1986, pp. 40-44.

[FAG95] Faghri, A. and J. Hua, "Roadway Seasonal Classification Using Neural Networks," *Journal of Computing in Civil Engineering*, No. 3, Vol. 9, July 1995, pp. 209-215.

[FAG96] Faghri, A., M. Glaubitz, and J. Parameswaran, "Development of Integrated Traffic Monitoring System for Delaware," *Transportation Research Record 1536*, Transportation Research Board, National Research Council, Washington, D.C., 1996, pp. 40-44.

[FLA93] Flaherty, J., "Cluster Analysis of Arizona Automatic Traffic Record Data," *Transportation Research Record 1410*, Transportation Research Board, National Research Council, Washington, D.C., 1993, pp. 93-99.

[FOT97] Fotheringham, A.S., M. Charlton, and C. Brunsdon, "Two Techniques for Exploring Non-Stationarity in Geographical Data," *Geographical Systems*, Vol. 4, pp. 59-82, 1997.

[FRA02] Fraley, C. and A.E. Raftery, *MCLUST: Software for Model-Based Clustering, Density Estimation and Discriminant Analysis*, Technical Report No. 415, Department of Statistics, University of Washington, Seattle, Washington, October 2002.

[FRA96] Fraley, C., *Algorithms for Model-Based Gaussian Hierarchical Clustering, Technical Report, No. 311*, Department of Statistics, University of Washington, Seattle, Washington, October 1996.

[FRA98] Fraley, C. and A.E. Raftery, "How Many Clusters? Which Clustering Method? Answers Via Model Based Cluster Analysis," *The Computer Journal*, Vol. 41, No. 8, 1998, pp. 578-588.

[FRE01] French, J.L., W. Iskander, and M. Jaraiedi, *Traffic Volume Related Research*, Final Report, prepared for Pennsylvania Department of Transportation, prepared by Pennsylvania Transportation Institute, Morgantown, West Virginia, May 2001.

[GRA98] Granato, S, "The Impact of Factoring Traffic Counts for Daily and Monthly Variation in Reducing Sample Counting Error," *Proceeding of Crossroads 2000*, Ames, Iowa, August 1998, pp. 122-125.

[GRI96] Griffith, D.A., "Spatial Autocorrelation and Eigenfunctions of the Geographic Weights Matrix Accompanying Geo-Referenced Data," *Canadian Geographer*, Vol. 40, No. 4, pp. 351-367, 1996.

[HAL84] Hallenbeck, M.E. and L.A. Bowman, *Development of a Statewide Traffic Counting Program Based on the Highway Performance Monitoring System*, Final Report, Prepared for Federal Highway Administration, March 1984.

[HPM00] *Highway Performance Monitoring System Field Manual*, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C., December 2000.

[JAN97] Jang, J., C. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, Prentice-Hall, Inc., Upper Saddle River, New Jersey, 1997.

[JAN98] Janikow, C.Z., "Fuzzy Decision Trees: Issues and Methods," *IEEE Transactions on Systems, Man, and Cybernetics*, Part B, Vol. 28, No. 1, February 1998, pp. 1-14.

[JOH02] Johnson, R.A. and D.W. Wichern, *Applied Multivariate Statistical Analysis*, Fifth Edition, Prentice-Hall, Inc., Upper Saddle River, New Jersey, 2002.

[KAM02] Kamvar, S.D., D. Klein, and C.D. Manning, "Interpreting and Extending Classical Agglomerative Clustering Algorithms Using A Model-Based Approach," *Proceedings of the 19th International Conference on Machine Learning*, 2002, pp. 283-290.

[KOP89] Kopanezou, H. and T. Trivellas, "A Time Series Model for Daily Traffic Volume Forecasting", *Proceedings of the Fourth International Conference on Civil and Structural Engineering Computing*, London, England, 1989, pp. 295-300.

[LAM00] Lam, W. and J. Xu, "Estimation of AADT from Short Period Counts in Hong Kong—A Comparison between Neural Network Method and Regression Analysis," *Journal of Advanced Transportation*, No. 2, Vol. 34, 2000, pp249-268.

[LI03] Li, M.T., F. Zhao, Y. Wu, and A. Gan, "Evaluation of Agglomerative Hierarchical Clustering Methods," Presented at the 82nd Transportation Research Board Annual Meeting, National Research Council, Washington, D.C, 2003.

[LIN00] Lingras, P., S.C. Sharma, P. Osborne, and I. Kalyar, "Traffic Volume Time-Series Analysis According to the Type of Road Use," *Journal of Computer-Aided Civil and Infrastructure Engineering*, No. 5, Vol. 15, 2000, pp. 365-373.

[LIN01] Lingras, P., "Statistical and Genetic Algorithms Classification of Highways," *Journal of Transportation Engineering*, No. 3, Vol. 127, 2001, pp. 237-243.

[LIN95] Lingras, P., "Classifying Highways: Hierarchical Grouping versus Kohonen Neural Networks", *Journal of Transportation Engineering*, No. 4, Vol. 121, 1995, pp. 364-368.

[MCD99] McDonald, S.M., *Prototype Demonstration of a Geographic Information System Application for the Seasonal Analysis of Traffic Data, Development of Seasonal Factors and Seasonal Adjustment of Roadways*, Final Report, Prepared for Federal Highway Administration, August 1999.

[MAS89] Massart, L.D. and L. Kaufman, *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*, Reprint, Robert E. Krieger Publishing Company, Malabar, Florida, 1989.

[OLA03] Olaru, C. and L. Wehenkel, "A Complete Fuzzy Decision Tree Technique," *Fuzzy Sets and Systems*, Vol. 138, 2003, pp. 221-254.

[PEN01] Peng, Y.H. and P.A. Flach, "Soft Discretization to Enhance the Continuous Decision Tree Induction," in Proceedings of ECML/PKDD-2001 Workshop IDDM-2001, Freiburg, Germany, 2001.

[PTFH02] *Project Traffic Forecasting Handbook*, Draft, Florida Department of Transportation, Tallahassee, Florida, May 2002.

[QUI86] Quinlan, J.R., "Induction of Decision Trees," *Machine Learning*, Vol. 1, 1986, pp. 81-106.

[RIT86] Ritchie, S.G., "A Statistical Approach to Statewide Traffic Counting", *Transportation Research Record 1090*, Transportation Research Board, National Research Council, Washington D.C., 1986, pp 14-21.

[SAK02] Sakawa, M., *Genetic Algorithms and Fuzzy Multiobjective Optimization*, Kluwer Academic Publishers, Norwell, Massachusetts, 2002.

[SAS99] *SAS-OnlineDOC*, Version 8, SAS Institute Inc., Cary, North Carolina, 1999.

[SEA00] Seaver, W.L., A. chatterjee, and M.L. Seaver, "Estimation of Traffic Volume on Rural Local Roads," *Transportation Research Record 1719*, Transportation Research Board, National Research Council, Washington D.C., 2000, pp 121-128.

[SHA00] Sharma, S.C., P. Lingras, G.X. Liu, and F. Xu, "Estimation of Annual Average Daily Traffic on Low-Volume Roads—Factor Approach Versus Neural Networks," *Transportation Research Record 1719*, Transportation Research Board, National Research Council, Washington D.C., 2000, pp 103-111.

[SHA01] Sharma, S.C., P. Lingras, F. Xu, and P. Kilburn, "Application of Neural Networks to Estimate AADT on Low-Volume Roads", *Journal of Transportation Engineering*, ASCE, Vol. 127, No. 5, 2001, pp 426-432.

[SHA81] Sharma, S.C. and A. Werner, "Improved Method of Grouping Provincewide Permanent Traffic Counters," *Transportation Research Record 815*, Transportation Research Board, National Research Council, Washington, D.C., 1981, pp. 12-18.

[SHA83] Sharma, S.C., "Improved Classification of Canadian Primary Highways According to Type of Road Use," *Canadian Journal of Civil Engineering*, No. 3, Vol. 10, 1983, pp. 497-509.

[SHA86] Sharma, S.C., P.J. Lingras, M.U. Hassan, and N.A.S. Murthy, "Road Classification According to Driver Population," *Transportation Research Record 1090*, Transportation Research Board, National Research Council, Washington, D.C., 1986, pp. 61-69.

[SHA93] Sharma, S.C. and R.R. Allipuram, "Duration and Frequency of Seasonal Traffic Counts," *Journal of Transportation Engineering*, No. 3, Vol. 119, 1993, pp. 344-359.

[SHA94] Sharma, S.C. and Y. Leng, "Seasonal Traffic Counts for a Precise Estimation of AADT," *ITE Journal*, September 1994, pp. 21-28.

[SHA96] Sharma, S.C., B.M. Gulati, and S.N. Rizak, "Statewide Traffic Volume Studies and Precision of AADT Estimates," *Journal of Transportation Engineering*, No. 6, Vol. 122, 1996, pp. 430-439.

[SHA99] Sharma, S.C., P. Lingras, F. Xu, and G.X. Liu, "Neural networks as Alternative to Traditional Factor Approach of Annual Average Daily Traffic Estimation from Traffic Counts", *Transportation Research Record 1660*, Transportation Research Board, National Research Council, Washington D.C., 1999, pp 24-31.

[SMI97] Smith, B.L. and M.J. Demtsky, "Traffic Flow Forecasting: Comparison of Modeling Approaches," *Journal of Transportation Engineering*, No.4, Vol. 123, 1997, pp. 261-266.

[STA97] Stamatiadis, N. and D.L. Allen, "Seasonal Factors Using Vehicle Classification Data," *Transportation Research Record 1593*, Transportation Research Board, National Research Council, Washington D.C., 1997, pp 23-28.

[TAN02] Tantrum, J., A. Murua, and W. Stuetzle, "Hierarchical Model-Based Clustering of Large Datasets through Fraction and Refractionation," *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 183-190.

[TMG01] *Traffic Monitoring Guide*, Office of Highway Policy Information, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C., May 2001.

[WEI96] Weinblatt, H., "Using Seasonal and Day-of-Week Factoring to Improve Estimates of Truck Vehicle Miles Traveled," *Transportation Research Record 1522*, Transportation Research Board, National Research Council, Washington D.C., 1996, pp 1-8.

[WIL98] Williams, B.M., P.K. Durvasula, and D.E. Brown, "Urban Freeway Traffic Flow Prediction: Application of Seasonal Autoregressive Integrated Moving Average and Exponential Smoothing Models," *Transportation Research Record 1644*, Transportation Research Board, National Research Council, Washington D.C., 1998, pp 132-141.

[WRI97] Wright, T., P.S. Hu, J. Young, and A. Lu, *Variability in Traffic Monitoring Data*, Final Summary Report, Prepared for U.S. Department of Energy, Prepared by Oak Ridge National Laboratory, Oak Ridge, Tennessee, August 1997.

[ZHA01] Zhao, F. and S. Chung, "Contributing Factors of Annual Average Daily Traffic in a Florida County," *Transportation Research Record 1769*, Transportation Research Board, National Research Council, Washington D.C., 2001, pp 113-122.