Seasonal Factor Assignment Based on the Similarity of Hourly Traffic Patterns and Influential Variables

Chenxi Lu* and Renfa Yang
Ningbo University of Technology / College of Transportation and Logistics, Ningbo, China
Email: {clu001@fiu.edu, yang0403@163.com}

Shanshan Yang and Fang Zhao
Florida International University / Department of Civil & Environmental Engineering, Miami, USA
Email: {syang003@fiu.edu, zhaof@fiu.edu}

***Abstract*—Average Annual Daily Traffic is typically estimated by applying seasonal factors (SFs) to short-term counts. SFs are obtained from continuous count sites and assigned to short-term count sites. This assignment procedure is usually empirical and subjective. Some previous studies have attempted to establish relationships between SFs and influential variables to provide an objective and data-driven alternative for SF assignment. However, in rural areas, SFs are difficult to model due to low land use intensity and, sometimes, significant through traffic. This paper presents a study of relationships between monthly SFs and hourly traffic patterns, land use, and other variables, using data from 116 continuous counters in rural areas throughout Florida. It is found that hourly traffic patterns are related to traffic seasonality and can be used to improve the modeling of influential variables that affect SF. The influential variables are then used for seasonal factor assignment and estimation. The proposed method achieved an average error of four percent, with 95 percent of the estimated monthly SFs having an error of no more than ten percent.**

***Index Terms*—Hourly traffic pattern, Seasonal factors, Land use, Spatial analysis, Regression analysis**

I. Introduction

Average Annual Daily Traffic (AADT) is an important traffic measure used in planning, traffic operations, safety analysis, pavement design, funding allocation, and other transportation applications. Transportation agencies in the U.S. and Canada expend significant resources to maintain their traffic monitoring systems, which require the collection of traffic data such as volume and classification counts. The *Traffic Monitoring Guide* (TMG)[1] requires that all agencies have a continuous and a short-term traffic counting program for federal reporting purposes. The agencies need to ensure that enough continuous data are collected to allow the calculation of AADT and seasonal adjustment factors. The TMG recommends a short-term count data collection program that covers an entire system on a six-year cycle, or covers a highway performance monitoring system (HPMS) sample and universal sections on a three-year cycle. Continuous counts are usually recorded by automatic traffic recorders. Short-term counts, also known as coverage counts, are usually conducted at selected sites for a period of 24 to 72 hours. To estimate AADT based on short-term counts, seasonal factors (SFs), which are obtained from continuous count sites, are assigned to a short-term count. The difficulty in this process lies in the fact that there is a lack of understanding as to what causes the differences in seasonal traffic variations, and how to determine if a continuous count site and a short-term count site share the same seasonal traffic pattern, and thus, the same SFs. Consequently, SF assignment procedures rely, to a large degree, on an analyst's experience and judgment, which may be prone to bias. This subjective SF assignment is likely to lead to inaccurate AADT estimates, which will affect the soundness of decisions made based on such AADT estimates.

The Florida Department of Transportation (FDOT) collects traffic data from about 300 telemetry traffic monitoring sites (TTMSs) located throughout the state. These TTMSs are continuous counters and their true AADT and SFs are available. There are more than 7,000 portable traffic monitoring sites (PTMSs), where only a short-term (for example, 72-hour) traffic count is conducted once or a few times a year. To estimate AADT for these PTMSs, the FDOT classifies the 300 TTMSs into 178 SF categories. The SF for a specific category is obtained by averaging the SFs from the TTMSs in the group. A PTMS is assigned a SF based on the consideration of roadway function classification, spatial proximity between the PTMS and nearby TTMSs, and the judgment of the analyst. The FDOT, however, desires a more data-driven and objective method that could improve the accuracy of the SF estimation for short-term count sites.

In a study conducted by Zhao *et al.*[2], land use and other variables that were considered to be influential to

doi:10.4304/jsw.8.4.868-877

SFs were explored for modeling SFs. It was found that in rural areas, SFs were difficult to model due to low land use intensity and, sometimes, significant through traffic. In this paper, a method for estimating SFs for short-term count sites on rural roads in Florida is proposed. This method jointly considers hourly traffic patterns, land use, and other variables. The TTMSs on rural roads are divided into two groups depending on whether their hourly traffic patterns exhibit commuting characteristics or not. Regression analysis is separately performed for the two groups to identify influential variables that may affect their SFs. The results indicate that considering hourly traffic patterns helps to significantly improve the models.

An assignment method is developed based on the influential variables identified in the regression analyses. This approach is based on the assumption that if variables identified in the regression analysis do contribute to seasonal traffic variations, they may be used to directly link a short-term count station, or PTMS, to a TTMS if the two share similar variable values. This method was tested with available TTMSs data. The test results show that the errors of estimated SFs for rural roads are, on average, four percent. The method developed here for rural roads was also applied to urban areas, which improves the SF assignment accuracy with a mean absolute percentage error of five percent[3].

In the remainder of this paper, the classification of hourly traffic patterns and regression analysis are described. The model results are then discussed and used for SF assignment.

II. Literature Review

Several past studies estimated AADT using independent variables and statistical methods, such as ordinary least square (OLS) regression[4,5,6], geographical weighted regression[7], and principal component analysis[8]. Independent variables used in these studies include roadway functional class, number of lanes, land use and socio-economic characteristics, *etc.* These methods do not require traffic data collection or SFs. However, the accuracy of these prediction methods is not high.

A commonly used approach for AADT estimation is to convert short-term counts to AADT either directly or indirectly. In direct conversion, regression is used to match short-term counters to continuous counters based on variations in traffic volume[9,10]. According to Robichaud[9], traffic data of at least eight days over three seasons are necessary to estimate AADT volumes with an accuracy of ±10 percent. Sharma *et al.*[10] employed an artificial neural network (ANN) method to estimate AADT from hourly traffic volume of a 48-hour count as input. They reported an estimation error of 7.9 percent, higher than the traditional factor method.

The most commonly applied approach in the U.S. is indirect conversion, i.e., factor-based method. Using this method, short-term traffic volumes (mostly 24- to 72-hour count) are usually collected once every two or three years. A short-term count is converted into AADT by assigning an SF to it. Data collection effort is minimal, but professional judgment and significant local knowledge are required for reasonable SF assignment. Sharma *et al.*[11] indicates that the accuracy is more sensitive to the correctness of SF assignment than the duration of short-term counts. The *Traffic Monitoring Guide*[1] provides recommendation with regards to applying SFs to short-term counts, however, local agencies are responsible for their own procedure.

Davis *et al.*[12] introduced the Bayesian assignment method for assigning SFs to short-term count sites. This method is mainly based on the matching of the monthly and day-of-week traffic variation pattern between short-term count sites and SF groups. The drawback is the need for monthly and weekly traffic variation information for short-term count sites, which requires significant data collection. Tsapakis[13] described a SF assignment approach based on statistical similarities in daily average traffic (ADT) and hourly traffic volumes between a short-period count and factor groups, which resulted in a 52 percent improvement in terms of mean absolute error over the method of assigning SFs based roadway function class.

Roadway functional classification (such as rural, urban, recreational, interstate, and collector) and locations have also been recognized as possible influential factors of seasonality[14,15,16]. Sharma[14] and Sharma *et al.*[15] proposed a method to classify rural roads based on trip purpose and trip length information collected from past origin-destination (OD) surveys by the Ministry of Transportation in Alberta, Canada. Based on the daily traffic patterns, five predominant road uses were identified[14]: commuter, commuter-recreational, commuter-recreational-tourist, tourist, and highly recreational. Three typical hourly traffic patterns were also identified: commuter, partially commuter, and non-commuter. Cumulative trip length distribution information, which was obtained from external station surveys, was used to classify roads for serving mainly regional, interregional, or long-distance travel. OD surveys, however, are expensive to conduct and are often economically infeasible, except in cases of corridor or intercity travel studies.

Other factors, such as demographics, socioeconomics, and land use types, have also been investigated as possibly explanations of seasonal traffic variations[2,3,17,18]. These factors, if understood and quantified, may aid in the assignment of SFs from one or more TTMSs to a PTMS, which may potentially reduce the data collection effort required and improve the accuracy of AADT estimations.

Believing that there is a connection between land use and seasonal traffic variations, Li *et al.*[17] employed a regression method to model the relationship between SFs and land use variables with limited TTMSs data on urban roads in South Florida. A number of influential variables were identified, including the concentration of seasonal households and retired households with high income, hotel/motel population, and retail employment. As in direct conversion method, the regression model is not

powerful enough to estimate SF directly with the identified influential factors. Li *et al.*[19] developed a fuzzy decision tree to classify a count site based on the value of selected variables that were identified in regression analyses. However, to validate the assignment results requires the considerable effort of collecting additional monthly short-term counts. Due to a lack of data, no validation was performed.

Zhao *et al.*[7] also modeled the relationship between SFs and land use patterns in rural areas in northern Florida. Variables such as the functional classification for highways, seasonal households, agricultural employment, and truck factor were identified as potential explanatory variables. However, monthly variation in traffic is more significant for rural roads than urban and commuter routes[20]. Modeling SFs for rural areas were found to be more difficult than for urban areas, likely due to low land use intensities, a lack of a dominant land use, and significant through traffic. The regression models for rural areas had low R-squared values when compared to the urban models.

In the remainder of this paper, an overview of a methodology for modeling SFs for rural areas in Florida, and assigning SFs obtained from TTMSs to PTMSs, is first provided. A detailed description then follows.

III. Overview Of Methodology

In this study, the methodology consists of three major steps: (1) multiple-linear regression analyses to identify potentially influential factors that contribute to the variations in SFs, and (2) a similarity-based assignment method that identifies TTMSs that are likely share the same MSFs with a shout count site based on the factors identified in the regression analyses, (3) estimation of the MSFs for a short count site based on those of the TTMS that are determined to be similar in Step (2).

In the regression analyses, the dependent variables are the 12 monthly seasonal factors (MSFs), and the dependent variables reflect land use, demographics, socioeconomics, economics, roadway, and other variables. The purpose is to identify variables that are statistically significant to MSFs. These variables are later assumed to provide the connection between a TTMS and a short count site that share similar MSFs.

It has been previously mentioned that land use is more difficult to model in rural areas than in urban areas due to low density, irregular roadway network, and more significant through traffic. Sharma[14] and Sharma *et al.*[15] suggested that commuting trips have their distinct traffic seasonal variability and hourly variability pattern. Commuting traffic typically has a distinct hourly pattern. The difference between the patterns of commuting and non-commuting traffic may be the result of differences in their land use patterns. This means that hourly traffic pattern may be used in modeling the SFs by ensuring variables included in a regression analysis are relevant to more data points.

In this study, before regression analysis is performed to test variables that may be important to SFs, the TTMSs in the rural areas of Florida are classified into two groups, commuting and non-commuting, based on their hourly traffic patterns. A commuting traffic pattern is characterized by double peaks (DP) during a day, one in the morning and the other in the afternoon, while non-commuting traffic has a single peak (SP) during the day. Regression analysis is separately performed for TTMSs in the SP and DP groups to model relationships between SFs and land explanatory variables.

Once regression models were developed, an assignment method was developed based on the influential variables identified in the regression analyses. Assuming these variables reflect the underlying causes for the seasonal traffic variations, they may be used to directly link one count station to a TTMS based on the similarity between their variable values. A similarity metric was developed that consisted of the influential variables weighted by their partial R-square values. A similarity score can be computed for any given count site based on the values of the influential factors for this site and the similarity metric. This method is tested with the TTMS data by comparing the SFs estimated based on the similarity scores against the known SFs. Test results show that the errors of the estimated SFs for rural roads are average four percent. The method developed here for rural roads was also applied to urban areas, which improves the SF assignment accuracy with a mean absolute percentage error of five percent[3].

In the following sections, classification of hourly traffic patterns, regression analysis, assignment and estimation of SFs, and conclusions are presented.

IV. Seasonal Factor Data

The basic data used in this study are the monthly seasonal factors (MSFs) from 116 TTMSs in Florida rural areas for the year 2000. The definition of rural areas is from the 2000 census. The MSFs data were provided by the Florida Department of Transportation. Figure 1 plots the location, AADT, and functional class of the roads where the TTMSs were sited.

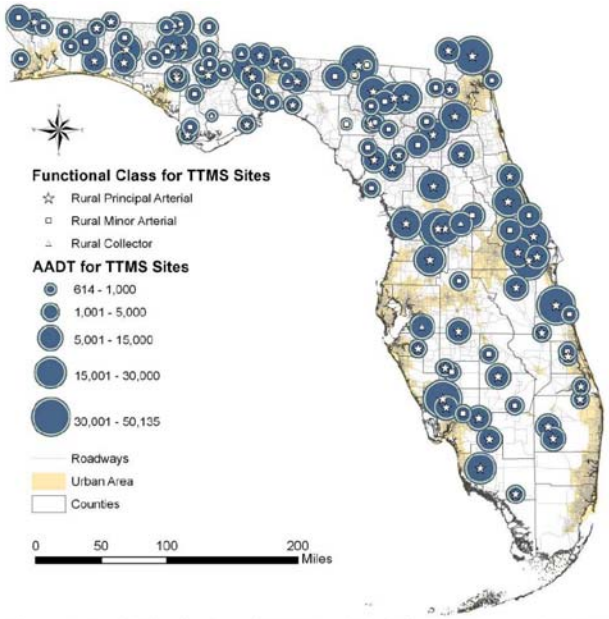

Figure 1 Spatial distribution of TTMSs and their function class and AADT

All of the 116 TTMSs have complete MSFs, including some (7.8 percent) that have been imputed due to missing data. A quality control and quality assurance procedure was followed to ensure the integrity and consistency of the data. All the imputed data have been reviewed and approved by the project team.

V. Classification Of Hourly Traffic Patterns

To study the hourly traffic patterns of the 116 TTMSs in rural Florida, Wednesday was chosen to represent a typical weekday. The hourly traffic volumes for all Wednesdays were extracted for each TTMS, which were then averaged to arrive at average weekday hourly volumes.

To determine if the hourly traffic pattern of a site exhibits a single peak (SP) or double peaks (DPs), the maximum and minimum of hourly volumes were examined. Figures 2(a) and 2(b) illustrate a single-peak traffic pattern and a double-peak traffic pattern, respectively. For both SP and DP patterns, Max_1 is defined as the maximum of hourly traffic volumes of T_i in the morning from hour 0 (0:00) to hour 10 (10:00), and Max_2 is the maximum of hourly traffic volumes in the afternoon from hour 15 (15:00) to hour 24 (24:00). *Min_midday* is the minimum of hourly volumes between the hour 10 (10:00) and hour 15 (15:00).

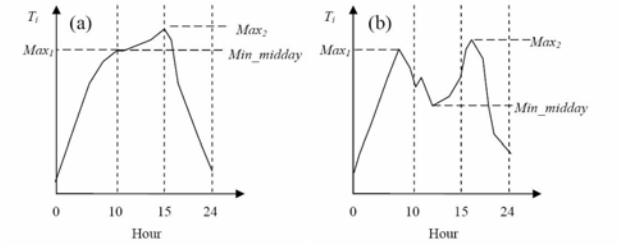

Figure 2. (a) Single-peak pattern and variables describing peaking characteristics; (b) Double-peak pattern and variables describing peaking characteristics.

A variable is defined to classify SP and DP pattern as follows:

$$Variation = \frac{\min(Max_1, Max_2) - Min\_midday}{Max(T_i)} \quad (1)$$

The variable *Variation* is guaranteed to be nonnegative by the definition of variables. If *Variation* is 0, then the low and high peaks are the same. This indicates that the traffic pattern has a single-peak, that is, a non-commuting pattern. If *Variation* is not zero, the traffic pattern may be considered to be a double peak (DP) pattern, which is characteristic of a commuter pattern. The larger *Variation* is, the more obvious the DP pattern is. However, a pattern will still be considered a SP pattern if *Variation* is not significant. The cutoff value of *Variation* need be defined to distinguish the SP and DP pattern. Four different criterion values (0.00, 0.05, 0.10, and 0.20) of *Variation* for classifying a TTMS into a SP or DP group are tested to investigate which one results in better models. The tests were performed on hourly data for Tuesday and Thursday that were averaged over the year for the same day. The same tests on Wednesday data were also conducted and resulted in the same classification groups. This suggests that the definition of hourly pattern is not sensitive to the day of data collection.

The purpose of averaging a full-year data is to ensure the smoothness of the data and that the data reflect the overall traffic pattern on an annual basis. However, short-term counts from PTMSs usually only cover one to three days, such as 72-hour duration. The traffic variation during such a short period may not be the same as that from the averaged data of an entire year. To verify whether SP or DP pattern based on annual average hourly traffic is similar to that of a short period count of 48 or 72 hours, the traffic patterns of all of the TTMSs for selected weekdays are examined. It is found that most of the SP or DP traffic patterns remain unchanged in different seasons, although volume variations are observed. Figure 3(a) shows the hourly traffic patterns at site 530050 in the months of January, April, July, and October. This site exhibits mostly consistent single-peak traffic patterns, even though the detailed hourly traffic pattern of each month varies. As an example, Figure 3(b) shows the DP hourly traffic patterns at site 500054 in the same four months, and the patterns are consistent.

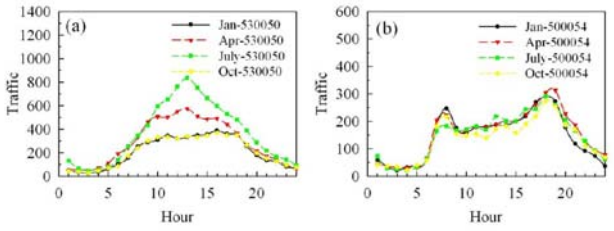

Figure 3. Hourly traffic pattern for two sample TTMS sites in different seasons (a) 530050 on US 231; (b) 500054.

When the cutoff criterion *Variation* = 0 is applied to annual averaged Wednesday hourly patterns, there are 33 count sites in the SP group and 83 in the DP group. When the cutoff criterion *Variation* = 0 is applied to short-term hourly patterns instead of the annual averaged patterns, 53% of the one-day hourly traffic patterns of the 33 sites in SP group are still classified as belonging to the SP group, while 89% of the one-day hourly traffic patterns of the 83 sites in the DP group are classified the same. Considering the hourly traffic pattern on any given weekday may have large variation or be less smooth than that averaged over all the same weekday over a year, the cutoff criterion *Variation* = 0.05 is suggested for use to classify short-term hourly patterns. Applying this criterion to the TTMSs resulted in 74% of the one-day hourly traffic patterns in the SP group remaining in the same group, and 80% of the one-day hourly traffic patterns in the DP group remaining unchanged.

VI. Multiple Linear Regression Analysis

Multiple linear regression analyses are conducted to identify potentially influential variables that contribute to the variations in the 12 monthly seasonal factors (MSFs) of the TTMSs. The stepwise selection method is applied, with the significance level set at 0.05 for a variable entering and staying in the model. The *t*-statistics and variance inflation factors (VIFs) are also checked for

each of the variables to ensure they are significant and to remove multicollinearity in the models.

Four different criterion values (0.00, 0.05, 0.10, and 0.20) of *Variation* for classifying a TTMS into a SP or DP group are tested to investigate which one results in better models. Applying each of the four criterion values, the TTMSs are classified into the SP and DP groups. Regression models are then developed. The dependent and independent variables are described in the next two sections.

A. Monthly Seasonal Factors as Dependent Variables

The dependent variables are the 12 MSFs, computed based on the weekday data. Because the FDOT traffic statistics program is mainly concerned with weekday traffic, a decision was made to remove traffic volume data from weekends, holidays, and special event days when, for example, hurricanes or large sport games occurred. The difference between the MSFs with or without the weekend data is found to be very small.

B. Definition of Independent Variables

Land use, from which travel activities are derived, is considered an important factor of both hourly and monthly traffic volume variations. Four categories of variables are considered in this research: roadway, demographics, employment, and geography.

The roadway variables include truck factor (T_F), roadway functional class (*FR*, *PA*, *MA* and *CO*), distance to the nearest interchange (*Interdist*), distance to the nearest metropolitan area (*D_Urban*), maximum ratio of population to the distance from a TTMS to a metropolitan area (*Dist1*), the inverse of the sum of ratios of urban population to the distance between a TTMS and an urban area (*Indexdist2*), and distance to the nearest beach (*Beachdist*). Variables *FR*, *PA*, *MA*, and *CO* are binary dummy variables. They indicate whether a road is a freeway (*FR*), principal arterial (*PA*), minor arterial (*MA*), or collector (*CO*).

The demographic variables include population density (*POPD*); percentage of seasonal households (*SHP*); percentage of retired households (*RETIRE*); percentages of population of different age groups: below 5 (*PPA5*), 6 to 17 (*PPA6_17*), 5 to 10 (*PPA5_10*), 11 to 13 (*PPA11_13*), 14 to 17 (*PPA14_17*), 18 to 21 (*PPA18_21*), 21 to 64 (*PPA22_64*), and 65 and up (*PPA65UP*); and density of population of different age groups: below 5 (*PDA5*), 6 to 17 (*PDA6_17*), 18 to 21 (*PDA18_21*), 21 to 64 (*PDA22_64*), and 65 and up (*PDA65UP*). The census data of year 2000 were used to calculate above variables.

The employment variables include employment density (*EMPD*); percentage of workers of different sectors: agriculture (*AgriP*), fishing and hunting (*FishP*), transportation (*TranP*), wholesale (*WholeP*), Retail (*RetailP*), restaurant (*ServP*), hotel and camp (*HotelP*), education (*EduP*), amusement and recreation (*RecServP*), and museum (*MuseumP*). Employment data for the year 2000 from the InfoUSA database were purchased by FDOT.

The climate changes significantly in Florida, from temperate in the North to subtropical in the South. This climate difference affects the seasonal activities of both the population and economy. For instance, South Florida attracts many visitors and welcomes the return of seasonal residents in its warm winter months. In contrast, summer in North Florida is the season for tourists and for outdoor recreations. For this reason, three climate zones are defined that divide the state into North, Central, and South Florida, and variables *SHP*, *HotlP*, *RtlP* and *MseumP* are defined for each of the climate zones by the prefix *N*, *C*, or *S* to indicate whether a TTMS is located in North, Central, or South Florida. These variables therefore become *NSHP*, *CSHP*, *SSHP*, *NHotlP*, *CHotlP*, *SHotlP*, *NRtlP*, *CRtlP*, *SRtlP*, *NMseumP*, *CMseumP*, and *SMseumP*, respectively.

Area-based variables, such as population, seasonal households, or employment, are measured using a buffer method. The buffer is created around a TTMS. However, because roadway spacing in rural areas is irregular, a uniform buffer size is inappropriate even for TTMSs on roads of the same functional classification. Therefore, a variable buffer method is used. Using this method, the distance between the road where a TTMS is located and the closest road that has the same functional classification is first computed using a geographic information system (GIS). A fixed percentage is then applied to this distance to determine the buffer size. Three percentages are tested with regression analysis: 25%, 50%, and 75%. Because 50% gives the best regression models, it is selected as the percentage used to compute the buffer size. For instance, if the distance between a TTMS and the next road with the same functional classification is eight miles, applying the 50% will give a buffer size of four miles. An upper limit of the buffer size of five miles is also imposed. The buffer area was restricted within rural areas if it overlapped with any urban areas.

C. Development of Multiple Linear Regression Models

Each of the four criterion values of *Variation* (0, 0.05, 0.1, and 0.2) led to two sets of models, one for the SP group and one for the DP group. Therefore, a total of eight sets of models were developed. Each set included 12 monthly models. Table 1 lists the adjusted R-squared values for the 12 MSF models for the four criterion values of *Variation*. The first column indicates the month. The second column provides the R-squared values for models that were calibrated with all 116 TTMSs without separating them into SP and DP groups. Columns 3 through 6 list the R-squared values for models corresponding to the four criterion values of *Variation*. The top half of table is for SP models and the bottom half of table is for DP models. The number of TTMSs used to develop the models is also given for each of group of models, which is the number of sites classified into the corresponding hourly pattern group by the different cutoff value of *variation* as indicated by *vari*.

TABLE 1.
COMPARISON OF ADJUSTED R-SQUARE FOR DIFFERENT MONTHLY SF MODELS.

	ALL	SP Models			
TTMSs	116	33	44	55	73
Month	*Vari* = 1.00	*Vari* = 0.00	*Vari* = 0.05	*Vari* = 0.10	*Vari* = 0.20
JAN	0.484	0.934	0.602	0.489	0.484
FEB	0.657	0.820	0.688	0.647	0.613
MAR	0.589	0.782	0.438	0.490	0.489
APR	0.297	0.635	0.175	0.179	0.180
MAY	0.218	0.489	0.491	0.488	0.420
JUN	0.452	0.621	0.531	0.562	0.447
JUL	0.501	0.900	0.511	0.484	0.465
AUG	0.520	0.641	0.472	0.606	0.571
SEP	0.531	0.812	0.625	0.453	0.549
OCT	0.238	0.370	0.381	0.439	0.469
NOV	0.347	0.764	0.463	0.359	0.465
DEC	0.408	0.607	0.596	0.490	0.471
		DP Models			
TTMSs		83	72	61	43
Month		*Vari* = 0.00	*Vari* = 0.05	*Vari* = 0.10	*Vari* = 0.20
JAN		0.573	0.541	0.587	0.600
FEB		0.606	0.654	0.621	0.593
MAR		0.552	0.514	0.680	0.777
APR		0.230	0.224	0.489	0.629
MAY		0.256	0.254	0.401	0.126
JUN		0.491	0.410	0.436	0.446
JUL		0.646	0.610	0.552	0.577
AUG		0.541	0.526	0.688	0.724
SEP		0.445	0.533	0.537	0.662
OCT		0.241	0.204	0.255	0.463
NOV		0.323	0.387	0.408	0.302
DEC		0.358	0.310	0.365	0.474

The model R-squared values suggest that separately modeling TTMSs with SP or DP patterns improve the explanatory power of the models. Although the improvement in model R-squared values for the DP models is not significant, the SP models have much higher R-squared values than the models without the SP and DP classifications. By comparing models based on the different criterion values of *Variation*, the SP models with criteria *Variation* = 0 have overall higher adjusted R-square values. This suggests that TTMSs pattern recognition may take the cutoff criterion *Variation* = 0.

The models for the SP group include about 30 variables. Among them, roadway variables, such as *Dist1*, *Interdist, Indexdist2*, and *TF* have relatively larger partial R-squared values. Population related variables, such as *SSHP*, *NSHP*, *PPA18_64*, and *PPA22_64*, also appear in the models and contribute noticeable partial R-squared values. Most of the employment variables, such as *SRtlP, NHotlP*, *ManuP*, *NMseumP*, *RcServP*, *FishP*, and *EdP*, enter the models, although some only appear in certain months and their partial R-Square values are not as significant as those of the roadway and population variables. The variables, their partial R-square values, and the model month are presented in Table 2. The signs of the coefficients of variables for each model month are presented in parentheses in the third column.

Table 2.
Variables in the SP Models (*Vari* = 0).

Variable	Partial R^2	Month
WholP	0.159	FEB (–)
WholP	0.060	JAN (–)
TranP	0.169	MAY (+)
TranP	0.074	JAN (–)
TranP	0.031	SEP (+)
TF	0.143	NOV (–)
TF	0.128	AUG (+)
TF	0.077	JUN (+)
TF	0.035	MAR (–)
SSHP	0.477	FEB (–)
SSHP	0.332	SEP (+)
SSHP	0.295	OCT (+)
SSHP	0.062	JAN (–)
SRtlP	0.233	AUG (+)
SRtlP	0.160	JUL (+)
SRtlP	0.114	OCT (+)
SRtlP	0.083	SEP (+)
SRtlP	0.078	MAR (–)
SRtlP	0.062	FEB (–)
SMseumP	0.033	MAR (–)
RestP	0.099	APR (–)
RcServP	0.088	DEC (+)
RcServP	0.042	NOV (–)
PPA65up	0.046	NOV (+)
PPA6_21	0.027	JUL (+)
PPA5_10	0.089	JUN (–)
PPA5_10	0.036	SEP (+)
PPA22_64	0.175	JUN (–)
PPA22_64	0.078	NOV (+)
PPA22_64	0.072	MAR (+)
PPA18_64	0.148	JAN (+)
PPA18_64	0.140	JUL (–)
PPA18_64	0.102	FEB (+)
PPA18_64	0.085	AUG (–)
PPA18_64	0.070	DEC (+)
PPA14_17	0.054	JUN (+)
PPA14_17	0.030	JAN (–)
PPA11_13	0.037	JUL (–)
NSHP	0.063	MAR (–)
NSHP	0.047	FEB (–)
NSHP	0.025	JUL (+)
NRtlP	0.052	APR (+)
NMseumP	0.164	MAR (–)
NHotlP	0.160	DEC (+)
NHotlP	0.081	SEP (+)
ManuP	0.263	JUL (+)
ManuP	0.079	MAY (–)
ManuP	0.077	APR (–)
ManuP	0.014	JAN (+)
Interdist	0.162	AUG (+)
Interdist	0.128	APR (–)
Interdist	0.010	MAY (–)
Interdist	0.080	JUL (+)

Variable	Partial R^2	Month
Indexdist2	0.118	NOV (+)
Indexdist2	0.094	JUL (–)
Indexdist2	0.041	JAN (+)
FishP	0.089	AUG (+)
FishP	0.057	SEP (+)
EdP	0.040	NOV (+)
EdP	0.028	JAN (+)
Dist1	0.387	MAR (–)
Dist1	0.345	JAN (+)
Dist1	0.339	DEC (–)
Dist1	0.285	JUN (+)
Dist1	0.274	APR (–)
Dist1	0.064	MAY (+)
CSHP	0.233	SEP (+)
CMseumP	0.073	APR (+)

Many variables help explain the seasonal variations of traffic. For instance, both *SSHP* and *NSHP* are significant, which suggests that seasonal households in South and North Florida influence traffic patterns in these regions. In winter months (such as January, February, or March), the coefficients of the *SSHP* (in January, and February) and *NSHP* variables (in February and March) are negative, suggesting that the seasonal residents who come down to Florida in the winter help increase traffic volumes. During the summer time, the coefficient of *SSHP* (in September and October) and *NSHP* (in July) variables are positive, suggesting that traffic is lighter because seasonal residents tend to stay outside of Florida in the summer and then return again in the winter. Variable *SRtlP* is more significant compared to *CRtlP* and *NRtlP*. This suggests that the seasonal pattern of traffic in the rural areas in South Florida is more affected by retail-related activities. The coefficient of the *SRtlP* is negative in winter months (in February and March) and positive in summer months (in July, August, September, and October), indicating that there is more shopping-related traffic in South Florida in the winter. Consequently, his may be attributed to the increase in the number of seasonal residents and tourists during this time.

Roadway variables, such as *Dist1*, *Indexdist2*, and *Interdist*, show significant correlation with SFs, but each has a different season during which they are important. For TTMSs closer to an interchange, traffic is lighter in April and May than in July and August. For TTMSs closer to a large metropolitan area, there is more traffic in March and April than in May and June.

However, inconsistent with conventional beliefs, the relationship between seasonal variation and roadway function class is not statistically strong based on the regression analyses, suggesting the current practice of SF assignment based on roadway function class needs to be reconsidered to determine its appropriate scope.

VII. Assignment And Estimation Of Seasonal Factors

A. Methodology for Measuring Similarity between Two Count Sites

Though the model is not power enough to predict SF based on the independent variables, the causal relation between SF and influential variables implies that MSFs are similar if PTMS shares similar characteristics with a TTMS. The goal of the assignment here is to identify a best matched TTMS or a number of best matched TTMSs for any given short-term count site based on their similarity scores.

The similarity score (or dissimilarity score), *S*, is calculated based on a selected set of variables that are identified in the regression analyses. To measure the similarity, the differences between the values of each of the variables for the two count stations are first computed. Recall that there are 12 regression equations in each model set, and that variables may appear repeatedly in different equations and are associated with different partial R^2 values. Hence, the partial R-squares of a variable from the monthly models are summed and used as weight to be applied to normalized differences. The sum of these weighted differences gives a score that measures weighted normalized differences for two count stations *i* and *j*, expressed as follows:

$$S_{ij} = \sum_{k=1}^{p} \frac{\left|V_{ki} - V_{kj}\right| \times SPR_k}{\max_k \{V_k\}}$$

where

S_{ij} = similarity score defined for count stations *i* and *j* ($i \neq j$),
p = number of influential variables,
V_{ki} = value of the *k*th variable in the 12-month models for count station *i*,
V_{kj} = value of the *k*th variable in the 12-month models for count station *j*,
SPR_k = sum of partial R^2 for the *k*th variable in the appeared months, and
$max(V_k)$ = maximum value for the variable V_k among all TTMSs.

Using the above definition, a similarity score can be computed for any pair of count stations. If multiple TTMSs are matched to a given count site, they may be ranked based on their similarity scores as the first best match with the lowest value of similarity score, second best match, and so on. SF may be assigned from a TTMS or TTMSs with the smaller value of similarity score.

B. Application of Similarity Scores to Rural TTMSs

To determine the effectiveness of the proposed similarity score, each of the 116 rural TTMSs was assumed to be a short-count site and was matched with other TTMSs. The evaluation of errors between an assumed short-count and matched TTMSs are conducted to arrive at an optimal method with an appropriate variable set and suitable TTMSs. The errors between matched sites are computed as follows:

$$e_{ij} = \frac{1}{12}\sum_{m=1}^{12}\left|\frac{MSF_{mi} - MSF_{mj}}{MSF_{mi}}\right| \quad (3)$$

where

e_{ij} = measure of difference between the monthly seasonal factors of count site i and j being compared,

MSF_{mi} = monthly seasonal factor for count site i for month m, and

MSF_{mj} = monthly seasonal factor for count site j for month m (may be mean of matched sites).

The complete variable set that has been identified in the regression analyses and is used to calculate similarity scores for the TTMSs in the SP group are *WholP, TranP, TF, SSHP, SRtlP, SMseumP, RestP, RcServP, PPA65up, PPA6_21, PPA5_10, PPA22_64, PPA18_64, PPA14_17, PPA11_13, NSHP, NRtlP, NMseumP, NHotlP, ManuP, Interdist, Indexdist2, FishP, EdP,* and *Dist1*. Due to a concern that too many influence variables are involved, a reduced variables set consisting of nine variables with the highest partial R-squared values instead of the original 25 is tested. This reduced variable set includes *TranP, TF, SSHP, SRtlP, PPA18_64, ManuP, Interdist, Indexdist2,* and *Dist1*.

For the DP group, the complete variable set includes 23 variables: *TF, SSHP, SRtlP, SHotlP, ServP, Rt_Low, Rt_High, RETIRE, RcServP, PPA65up, PPA22_64, PPA14_17, PPA11_13, NSHP, MineP, Indexdist2, FR, EdP, Dist1, CO, CHotlP, Beachdist,* and *AgriP*. A reduced set of variables is also tested, which include eight variables with the highest partial R-squared values: *TF, SSHP, SRtlP, Rt_High, RETIRE, NSHP, FR, Dist1,* and *Beachdist.*

Table 3 shows the average errors between MSFs estimated from matching TTMSs and true MSFs for all rural TTMSs in the SP and DP groups with both complete and reduced influence variable sets. The last row of the table gives the errors in the estimated MSFs that have been obtained by averaging the MSFs of the first two best matched sites.

TABLE 3.
AVERAGE ERRORS OF THE ASSIGNMENT RESULTS FOR RURAL AREA.

Best Match	SP Complete Variable Set	SP Reduced Variable Set	DP Complete Variable Set	DP Reduced Variable Set
1st	4.2%	4.1%	4.1%	4.5%
2nd	4.8%	5.1%	4.1%	4.4%
3rd	4.9%	5.5%	4.5%	4.1%
4th	5.6%	4.8%	4.3%	4.4%
5th	5.2%	5.6%	4.9%	4.7%
(1st + 2nd)/2	4.1%	4.2%	3.6%	3.9%

It can be seen from Table 3 that the reduced variables set produced assignment results with good accuracy. It is recommended to estimate MSFs for a count site by averaging the corresponding MSFs of the first two (or more) best matches. In addition to reduced errors, using two matched sites instead of a single one may also increase the reliability, and possibly decrease randomness, in the first best match.

VIII. EVALUATION OF PROPOSED MSF ESTIMATION METHOD

The accuracy of the MSF estimation method is evaluated by estimating the 12 MSFs for all of the 116 TTMSs, i.e., 116 × 12 = 1,392 MSFs and comparing them to the actual MSFs of the TTMSs. The absolute percentage error between the MSFs of two matched sites is used to evaluate the accuracy of the assignment method, as follows:

$$e_{mi} = \left|\frac{MSF^{a}_{mi} - MSF^{e}_{mi}}{MSF^{a}_{mi}}\right| \quad (4)$$

where

e_{mi} = absolute percent difference between the actual and estimated monthly seasonal factors of count site i for month m,

MSF^{a}_{mi} = actual monthly seasonal factor for count site i for month m, and

MSF^{e}_{mj} = estimated monthly seasonal factor for count site j for month m.

Table 4 shows the error frequency distribution of the estimated MSFs using the full variable set. The mean error is 3.7 percent, and the maximum error is 25 percent for one month. Over 95 percent of assigned MSFs have an error below 10 percent.

TABLE 4.
ERROR DISTRIBUTION OF ESTIMATED MSFS FOR RURAL TTMSS BASED ON THE FULL VARIABLE SET.

Error Range	Frequency of MSFs	Percentage of Frequency	Accum. Frequency of MSFs	Percentage of Accum. Frequency
[0%, 2%]	501	36.0%	501	36.0%
(2%, 4%]	380	27.3%	881	63.3%
(4%, 6%]	255	18.3%	1,136	81.6%
(6%, 8%]	122	8.8%	1,258	90.4%
(8%, 10%]	74	5.3%	1,332	95.7%
(10%, 12%]	28	2.0%	1,360	97.7%
(12%, 14%]	17	1.2%	1,377	98.9%
(14%, 16%]	2	0.1%	1,379	99.1%
(16%, 18%]	8	0.6%	1,387	99.6%
(18%, 20%]	2	0.1%	1,389	99.8%
(20%, 25%]	3	0.2%	1,392	100.0%

The error frequency distribution of the estimated MSFs using the reduced variable set is similar to the results obtained using the full variable set. The mean absolute error for all estimated MSFs is 4.0 percent, and the maximum error is 27 percent. Nearly 95 percent of the estimated MSFs have an error below 10 percent. The distribution of error frequency is depicted in Figure 4.

Please note that errors in MSFs directly translate into errors into AADT estimation. Because when ADT is converted into AADT, the SF applied is the weekly SF interpolated from the two adjacent MSFs, the errors in MSF estimates provide an upper bound of errors in the corresponding AADT estimates.

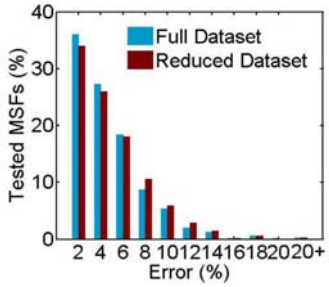

Figure 4. Error distribution of estimated MSFs for all rural TTMSs.

The results suggest that the proposed assignment method is able to achieve good accuracy, especially considering that traffic variation in rural areas is more significant than in urban areas and land use is more challenging to model.

IX. Conclusions

The literature has indicated that it is challenging to identify the relationship between monthly seasonal factors (MSFs) and influential variables in Florida rural areas. This research has shown that hourly traffic patterns are important and can be used to build better rural models for seasonal factor (SF) analysis. By classifying TTMSs into single-peak (SP) and double-peak (DP) groups, regression models that relate MSFs to various roadway and influential variables achieve higher adjusted R-squared values, especially when traffic is less dominated by commuting trips. This suggests that there exists an intrinsic connection between the hourly traffic pattern, the seasonal traffic pattern, and influential variables.

The regression analysis identified two sets of influential variables for TTMS with single- and double-peak weekday hourly traffic patterns. A similarity score is defined based on the influential variables, which is used to match two count sites based on their similarity in the values of the influential variables. The MSFs of a short-term count site can then be estimated by averaging the MSFs of the two ore more best matched TTMSs. One interesting observation is that contrary to conventional beliefs, roadway function classification did not turn out to be a strong statistical indicator for MSFs. This does not mean that function class should be completely disregarded. Especially for freeway sections, along which there are no major disjoint land uses, SFs on these sections may be similar because they reflect the seasonality of through traffic that is not affected by local economic activities.

Application of the proposed method can begin by classifying a PTMS as a single-peak or double-peak site based on hourly traffic volumes from short-term counts. Depending on the classification of the PTMS, the reduced variable set from the SP or DP regression models can then be used to compute the similarity score and match it with TTMSs that share similar roadway and land use characteristics. The seasonal factors of the best matching TTMSs may be selected to estimate the SFs for the PTMS.

The proposed method was tested by estimating MSFs for the 116 rural TTMSs and comparing them to the known MSFs. Of the 1392 estimated MSFs of the 116 test sites, 75 percent had an error of 6 percent or less, and 95 percent had an error within 10 percent. Note that 10 percent is the threshold considered to be acceptable by the FDOT when estimating MSFs.

The assignment method developed in this study offers at least three advantages. First, no additional TTMSs are required to validate the assignment results. This makes this approach more practical and less expensive when compared to, for example, a fuzzy decision tree. Second, a count site may be linked to multiple TTMSs. This provides the analyst with alternative TTMSs in case there is a sufficient basis to reject the best matching TTMS based on the selected variables. Third, this method can be tested with the same TTMSs that are used in the regression analysis. Although this is not to say that there is no need for independent testing using an entirely different set of data, this method allows the development of some understanding of how well the method works. Finally, this method has the potential to eliminate the need to conduct SF grouping.

The regression variables used in this study have been carefully selected to ensure that they capture rich information that may affect SFs while being readily available from either census or transportation planning data. Although land use and demographics may evolve over time, this evolution is usually slower in rural areas than urban areas and frequent model updates may not be needed. Therefore, model updates at intervals of five to ten years may be adequate. The models, however, need to be developed for specific regions if applied outside Florida.

Future research will be focused on further analysis of the variables identified in the regression models and exploring other traffic parameters that may bear on SFs.

Acknowledgements

The work is part of Programs Supported by Ningbo Natural Science Foundation under Grant 2012A610156. This research was also partly funded by the Research Center of the Florida Department of Transportation, with support from the FDOT gratefully acknowledged. The opinions, findings, and conclusions expressed in this paper are those of the authors and not necessarily those of the FDOT.

References

[1] USDOT (2001). *Traffic Monitoring Guide*. U.S. Department of Transportation and Federal Highway Administration, Washington, D.C.

[2] Zhao, F., Yang, S.S., and Lu, C.X. (2008). *Alternatives for Estimating Seasonal Factors on Rural and Urban Roads in*

Florida (Phase II). Final Report for Project BD015-17, Research Office, Florida Department of Transportation, Tallahassee, FL.

[3] Yang, S.S., Lu, C.X., Zhao, F., Reel, R. and O'Hara, J.D. (2009). "Estimation for Seasonal Factors of Similarity-Based Traffic for Urban Roads in Florida." *Transportation Research Record: Journal of the Transportation Research Board, No. 2121,* Transportation Research Board of the National Academies, Washington, D.C., pp. 74–80.

[4] Neveu, A. J. (1983). "Quick Response Procedures to Forecast Rural Traffic." *Transportation Research Record: Journal of the Transportation Research Board, No. 944,* Transportation Research Board of the National Academies, Washington, D.C., pp. 47–53.

[5] Mohamad, D., Sinha, K. C., Kuczec, T., & Scholer, C. F. (1998). Annual Average Daily Traffic Prediction Model for County Roads. *Transportation Research Record: Journal of the Transportation Research Board, No. 1617,* Transportation Research Board of the National Academies, Washington, D.C., pp. 69–77.

[6] Zhao, F. and S. Chung (2001). "Contributing Factors of Annual Average Daily Traffic in a Florida County." *Transportation Research Record: Journal of the Transportation Research Board, No. 1769,* Transportation Research Board of the National Academies, Washington, D.C., pp. 113–122.

[7] Zhao, F., & Park, N. (2004a). "Using Geographically Weighted Regression Models to Estimate Annual Average Daily Traffic." *Transportation Research Record: Journal of the Transportation Research Board, No. 1879,* Transportation Research Board of the National Academies, Washington, D.C., pp. 99–107.

[8] Seaver, W. L., Chatterjee, A., & Seaver, M. L. (2000). "Estimation of Traffic Volume on Rural Local Roads." *Transportation Research Record: Journal of the Transportation Research Board, No. 1719,* Transportation Research Board of the National Academies, Washington, D.C., pp. 121–128.

[9] Robichaud, K. and M. Gordon (2003). "Assessment of Data-Collection Techniques for Highway Agencies." *Transportation Research Record: Journal of the Transportation Research Board, No. 1855,* Transportation Research Board of the National Academies, Washington, D.C., pp. 129–135.

[10] Sharma, S. C., Lingras, P.j., Xu, F., & Liu, G. X. (1999). "Neural Networks as Alternative to Traditional Approach of Annual Average Daily Traffic Estimation from Traffic Counts." *Transportation Research Record: Journal of the Transportation Research Board, No. 1660*, Transportation Research Board of the National Academies, Washington, D.C., pp. 24-31.

[11] Sharma, S.C., Gulati, B.M., and S.N. Rizak. (1996). "Statewide Traffic Volume Studies and Precision of AADT Estimates." *Journal of Transportation Engineering*, 122 (6),pp. 430-439.

[12] Davis, G. A., and Guan, Y. (1996). "Bayesian Assignment of Coverage Count Locations to Factor Groups and Estimation of Mean Daily Traffic." *Transportation Research Record: Journal of the Transportation Research Board, No. 1542*, Transportation Research Board of the National Academies, Washington, D.C., pp. 30–37.

[13] Tsapakis, I. (2009). *Determination of Seasonal Adjustment Factors and Assignment of Short-Term Counts to Factor Groupings*. Ph.D dissertation, Department of Civil Engineering, University of Akron, Akron, OH.

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

[15] Sharma, S.C., Lingras, P.J., Hassan, M.U. and Murthy, N.A.S. (1986). "Road Classification According to Driver Population," *Transportation Research Record: Journal of the Transportation Research Board, No. 1090*, Transportation Research Board of the National Academies, Washington, D.C., pp. 61-69.

[16] Capparuccini, D.M., Faghri, A., Polus, A., and Suarez, R.E. (2008). "Fluctuation and Seasonality of Hourly Traffic and Accuracy of Design Hourly Volume Estimates." *Transportation Research Record: Journal of the Transportation Research Board, No. 2049,* Transportation Research Board of the National Academies, Washington, D.C., pp. 63–70.

[17] Li, M.T., Zhao, F. and Wu, Y. (2004). "Application of Regression Analysis for Estimating Seasonal Factors in Southeast Florida", *Transportation Research Record: Journal of the Transportation Research Board, No. 1870*, Transportation Research Board of the National Academies, Washington D.C., pp. 153-161.

[18] Zhao, F., Li, M.-T., and Chow, L.-F. (2004b). *Alternatives for Estimating Seasonal Factors on Rural and Urban Roads in Florida.* Final Report prepared for the Research Center, Florida Department of Transportation, Tallahassee, FL. Available online (http://www.dot.state.fl.us/research-center/Completed_Proj/Summary_PL /FDOT_BD015_03_rpt.pdf).

[19] Li, M.T., Zhao, F. and Chow, L. (2006). "Assignment of Seasonal Factor Categories to Urban Coverage Count Stations Using a Fuzzy Decision Tree", *J. Transp. Engrg.*, Vol. 132, No. 8, pp. 654-662.

[20] HCM (2000). *Highway Capacity Manual 2000*, Transportation Research Board, National Research Council, Washington, D.C.