Identifying Time-of-Day Breakpoints Based on Nonintrusive Data Collection Platforms

Rui Guo & Yu Zhang

Taylor & Francis
Taylor & Francis Group

Identifying Time-of-Day Breakpoints Based on Nonintrusive Data Collection Platforms

RUI GUO and YU ZHANG

Department of Civil and Environmental Engineering, University of South Florida, Tampa, Florida, USA

To ensure the effective operation of traffic signal systems, different signal timings should be designed to accommodate traffic pattern variations. One of the greatest challenges is the identification of appropriate time-of-day (TOD) breakpoints, where different signal timings could be implemented during the time periods between two consecutive breakpoints. This research presents an advanced cluster analysis aimed at identifying TOD breakpoints for coordinated, semiactuated modes when it is necessary for multiple intersection operations to be considered simultaneously. Different from previous studies, this proposed methodology considers the time of traffic occurring as one dimension in clustering and uses continuous traffic data obtained through innovative, nonintrusive data collection techniques, which significantly improve this method's performance. The operability of this proposed method is demonstrated in a case study of a corridor located in Tampa, FL. The traffic simulation results reported in this article reveal that this novel procedure performs better than existing TOD signal timing plans.

Keywords Cluster Analysis; Nonintrusive Data Collection; Time-of-Day Breakpoints

INTRODUCTION

Traffic signal systems serve as one of the most powerful control tools in improving the efficiency of surface transportation travel. The basic premise behind traffic signal control is the development of signal timing plans that are best suited for expected traffic conditions for particular dates or times. Considering daily traffic pattern variations, it is necessary to determine appropriate time-of-day (TOD) breakpoints, where different signal timings could be implemented during the time periods between two consecutive breakpoints. Transition costs will occur when changing timing plans or entering into coordinated timing plans, because it takes time for controllers to operate transition algorithms and shift local offset reference points (Koonce, Rodegerdts, Lee et al., 2008). Therefore, the determination of TOD breakpoints needs to balance the efficiency of signal timing and the consequent transition costs.

The experience of traffic engineers and an imprecise analysis of traffic volume data usually determine current day plan schedules. This traditional method contains many subjective factors and can easily lead to unreasonable divisions. In many cases, the signal system does not operate very efficiently. Though optimization tools are available to assist traffic engineers in developing timing plans, few tools exist to help them determine appropriate TOD breakpoints. This deficiency has motivated researchers to develop an improved procedure for assisting in the determination of TOD breakpoints for traffic signal systems.

A significant opportunity to address the need for more efficient signal timing plans lies in the fact that advanced communications, electronics, and information technologies, commonly referred to as intelligent transportation systems (ITS), have been widely deployed and plays an increasingly important role in improving the efficiency, safety, and reliability of transportation systems (Smith & Venkatanarayana, 2005). ITS has the ability to record large quantities of traffic condition data collected by existing system detectors or nonintrusive data collection platforms. To illustrate the potential of advanced data collection techniques, a cluster analysis-based procedure is developed in this study to determine TOD breakpoints for coordinated, semiactuated traffic systems using continuous traffic data samples obtained from an advanced ITS data collection platform.

Based on a thorough literature review, none of the previous cluster analysis-based studies have explicitly incorporated

Address correspondence to Yu Zhang, Department of Civil and Environmental Engineering, University of South Florida, 4202 E. Fowler Ave, ENB118, Tampa, FL 33620, USA. E-mail: yuzhang@usf.edu

the time of traffic occurring as one dimension in a clustering analysis to find the optimal TOD breakpoints (Hauser & Scherer, 2001; Park et al., 2003; Smith et al., 2002; Wang et al., 2005). Nevertheless, time factors should not be ignored when considering the inherent continuous nature of traffic volumes throughout the day. In this article, the time feature is incorporated into an advanced clustering analysis to identify the TOD breakpoints for traffic signal timing plans. The remainder of this article is organized as follows. The next section reviews the literature relative to TOD breakpoint determination and provides a background on current traffic signal timing practices and procedures. Afterward, the methodologies based on the application of this advanced cluster analysis procedure are proposed. Next, the ways of collecting different types of data are introduced before discussing the details of data analysis. Finally, the results of a case study conducted to test this procedure are presented and deliberated along with a comparison of an existing day plan schedule.

LITERATURE REVIEW

The typical approach used to identify intervals for TOD signal plans is to plot aggregate traffic volumes over the course of a day for representative sample intersections. The significant changes in traffic volume, which indicate a need for different timing plans, are then manually determined based on engineering judgment (Koonce et al., 2008; Smith et al., 2002). These intervals rely heavily on the existing traffic conditions at typical intersections throughout the arterial. As Abbas (2005) pointed out, unless traffic patterns change at certain times of the day and remain constant until the next change—which is highly unlikely—it is very difficult to determine the optimal breakpoints. However, considering the intrinsic variation of traffic volumes, several researchers have proposed methodologies for selecting TOD breakpoints and have demonstrated the benefits of doing so in traffic operations (Abbas, 2005; Hauser & Scherer, 2001; Park et al., 2003; Smith et al., 2002; Wang et al., 2005). The majority of these studies can be categorized into two groups: a heuristic search algorithm-based approach and a cluster analysis-based approach.

Searching for the best TOD breakpoints can be constructed as a mathematical optimization problem. Park et al. (2003) adopted a genetic algorithm (GA) into an intervals identification process to find the best TOD breakpoints. Good results were obtained from their study, but the results needed to be modified artificially because of the premature convergence of the GA (Yang et al., 2006). In a follow-up study, they presented a developed procedure for determining optimal breakpoints on TOD-based coordinated actuated traffic signal operations that used a feature vector of optimal cycle length per time interval instead of traffic volume itself (Park & Lee, 2008; Lee et al., 2011). Nevertheless, their studies were only conducted for hypothetical arterial networks. The performance of their GA-based optimization needs to be verified in real traffic situations. At the 84th Transportation Research Board (TRB) Annual Meeting, Abbas (2005) introduced a multi-objective evolutionary algorithm, nondominated sorting GA, with degree of detachment (DOD). However, this algorithm does not ensure the global optimal solutions for all cases. Similarly, Yang et al. (2006) developed an artificial immune systems (AIS)-based data analysis algorithm. At a certain level, the algorithm reduces redundant information of sample traffic data and overcomes the irrationality of the artificial method. However, there are too many adjustable parameters in the algorithm and it is very difficult to select the proper parameters to illustrate objective functions or rationally add constraints.

The other approach for identifying TOD breakpoints is to apply a cluster analysis. Hauser and Scherer (2001) and Smith et al. (2002) studied the possibility of period division by employing a hierarchical cluster analysis to identify optimal TOD breakpoints. Their studies demonstrated the feasibility of cluster analysis through a single intersection case study. However, traffic data falling into two time intervals far away from each other along the time axis were clustered into one group. This would cause some clusters to jump around adjacent intervals and signal timing plans would have to be changed frequently, leading to high transition costs. In 2005, Wang et al. proposed a nonhierarchical cluster analysis called K-means clustering. This technique provides several benefits, including simplicity, reduced data storage, and user-defined clusters (Lee et al., 2011). Their study applied the K-means clustering algorithm to identify optimal TOD breakpoints, which was based on a very limited data resource. Therefore, the treatment of the frequent transitions issue remained unresolved. Xia and Chen (2007) also used a data-clustering technology to define flow phases based on site-specific historical traffic data obtained through detectors. However, the imprecise method for determining the number of clusters was inadequate. Ratrout (2010, 2011) demonstrated his most recent research results of determining optimum TOD breakpoints based on the K-means technique. Concerned by cyclic traffic along arterials with pretimed controllers, he considered 24-hour volumes twice to form continuous traffic data spreading over 48 hours. His research, though promising, did not take into account the coordination effect of intersections because he focused on developing timing plans for pretimed signal controllers.

Although the cluster analysis-based studies have shown the feasibility of identifying TOD breakpoints, none of them has explicitly incorporated the time of traffic occurring as a feature in a clustering analysis. However, treating time-series data as an unordered collection of events and ignoring its temporal information leads to excessive transitions, scattered outliers, and consequent high operational costs and traffic delay. Recent theoretical studies in cluster analysis have demonstrated the effectiveness of using background information at the instance level to create must-link and cannot-link constraints (Everitt et al., 2011). A must-link constraint enforces two instances to be included in the same cluster, while a cannot-link constraint enforces the two instances not to be placed in the same cluster.

In our study, time is added as an additional dimension in the cluster analysis, which implicitly defines must-link and cannot-link constraints for this particular research problem.

In summary, much work on TOD breakpoint determination has been undertaken. Although the effectiveness of previous techniques has been demonstrated, due to insufficient methodologies for coordinated semi-actuated modes, there are still obvious shortcomings. The major challenges of applying a cluster analysis-based approach to identify TOD breakpoints include accurate traffic data resources, the need for adequate considerations of time–space traffic data features, and the need for a systematic framework that can be easily implemented. Our study fills in the gaps and tackles the challenges by proposing an improved cluster analysis and a framework for practical implementation.

METHODOLOGY AND RESEARCH APPROACH

Methodology

Traffic patterns show significant variations throughout the day. To make signal timing controllers work effectively, different signal timing plans should be designed in appropriate intervals where the traffic in one interval is relatively stationary. Cluster analysis, which addresses the problem of data segmentation, belongs to unsupervised learning methods since there is no knowledge of "preferred" clusters (Duda et al., 2001). It is a class of techniques used to classify a set of data into groups that are relatively homogeneous within themselves and heterogeneous between each other on the basis of a defined set of variables. Important issues, including cluster elements, distance measures, number of clusters, clustering algorithms, and validation of the analysis, must be considered to conduct a successful cluster analysis.

Cluster Elements

Regardless of the needs of a particular application, data samples are usually multidimensional. Thus, an essential step in clustering is to select the proper cluster elements that define the system states. Traffic volumes (or rate of traffic flow, measured in vehicles/hour) from intersection approaches are the most commonly used cluster elements in previous TOD breakpoint studies (Hauser & Scherer, 2001; Lee et al., 2011; Park et al., 2003; Park & Lee, 2005; Ratrout, 2010, 2011; Smith et al., 2002; Wang et al., 2005). As mentioned earlier in this article, due to the inherent continuous nature of daily traffic volumes, the time that traffic occurs should be considered as an additional dimension. Also, for coordinated semiactuated modes, the main road traffic volumes at major intersections (within the study corridor) should be considered simultaneously. Therefore, the system states in our study are defined as follows, assuming there are M representative segments through

Table 1   Format of data used for advanced cluster analysis.

Time-of-day	Cluster elements						
1	x_{11}	x_{12}	$\cdots$	x_{1m}	$\cdots$	x_{1M}	$x_{1(M+1)}$
2	x_{21}	x_{22}	$\cdots$	x_{2m}	$\cdots$	x_{2M}	$x_{2(M+1)}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
t	x_{t1}	x_{t2}	$\cdots$	x_{tm}	$\cdots$	x_{tM}	$x_{t(M+1)}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
T	x_{T1}	x_{T2}	$\cdots$	x_{Tm}	$\cdots$	x_{TM}	$x_{T(M+1)}$

the entire corridor and daily traffic is recorded into T small time intervals:

$$X_t = (x_{t1}, x_{t2}, \cdots, x_{tm}, \cdots, x_{tM}, x_{t(M+1)})$$

where X_t is the system state at time t, $t = 1, 2, \cdots, T$, and X_{ip} is the traffic volume in segment M at time t, p $= 1, 2, \cdots, M$, and $x_{t(m+1)}$ is the time variable.

The data used for this study will have the format as shown in Table 1.

To deal with the differences in scale between traffic volumes and time variables at different intersections, the cluster elements should be properly standardized (Hauser & Scherer, 2001; Ratrout, 2010, 2011). This process, which uses Eq. 1, is performed prior to the cluster analysis so as to make original traffic data dimensionless.

$$x'_{tm} = \frac{x_{tm} - \bar{x}_m}{s_m} \quad (t = 1, 2, \cdots, T; \, m = 1, 2, \cdots, M, M+1) \quad (1)$$

where x_{tm}, $\bar{x}_m$ and s_m represent original, average, and standard deviation of traffic volumes or time variables, respectively, for any particular observation.

In cluster analysis, cluster elements are grouped according to their similarities, or more specifically, the distances between them. Therefore, the smaller the distances between the elements, the more similar they are and the more likely it is that they belong to the same cluster (Smith et al., 2002). For our study, squared Euclidean distance, as shown in Eq. 2, is implemented for calculating the distance between clusters:

$$d_{ij}^2 = \sum_{m=1}^{M+1} (x_{im} - x_{jm})^2 \quad (i, j \in [1, T]; \, m \in [1, M+1]) \quad (2)$$

Where d_{ij}^2 is the squared Euclidean distance between state element i and j; x_{im} is the mth element in state i and x_{jm} is the mth element in state j.

Number of Clusters

Determining the optimal number of clusters, which directly affects the day plan schedule performance, is one of the most critical steps in the cluster analysis. An insufficient number of clusters would muddle up different traffic patterns, while

Table 2 Proposed and existing breakpoints and TOD intervals.

Proposed breakpoints and TOD intervals					
Free-flow	Pre-AM	AM peak	Mid-day	PM peak	PM off-peak
19:45–6:00	6:00–6:45	6:45–9:00	9:00–14:00	14:00–18:30	18:30–19:45
Existing breakpoints and TOD intervals					
Free-flow	AM peak	Mid-day	PM peak	PM off-peak	Evening
22:00–6:00	6:00–9:30	9:30–14:00	14:00–15:00	15:00–19:00	19:00–22:00

too many clusters would be prohibitively costly and impractical. Two criteria, the silhouette measure and the gap-statistic measure, are used in this study to identify the proper number of clusters. By comparing the outputs of different validation methodologies, the reliability of the results by showing whether or not the estimations crucially depend on a particular criterion can be checked.

The silhouette measure operates on the basis of the dissimilarity matrix suggested by Rousseeuw (1987). It is a popular method used to specify, in advance, the optimal number of clusters to be used. Using this approach, each cluster can be represented by a so-called silhouette, which is based on the comparison of its tightness and separation. The average silhouette width, defined by Eq. 3, is used to decide how good the number of selected clusters is:

$$s(i) = \frac{\min\{D_{ij}, j \in C_{-i}\} - D_{iC_i}}{\max(\min\{D_{ij}, j \in C_{-i}\}, D_{iC_i})} \quad (3)$$

where C_{-i} denotes cluster labels that do not include element i as a member; C_i denotes the cluster label that includes element i; D_{ij} is the averaged distance between element i and all elements in cluster C_{-i} (in other clusters); $\min\{D_{ij}, j \in C_{-i}\}$ is the minimum of average dissimilarity *of i* to all elements in another cluster (in the closest cluster); and D_{iC_i} is the averaged distance between element i and all elements in cluster C_i (in the same cluster).

Based on these individual data, the overall average silhouette width, denoted as the silhouette coefficient (SC), is simply the average of the $s(i)$ for all elements in the data set as shown in Eq. 4:

$$SC = \frac{1}{N} \sum_{i=1}^{N} \frac{\min\{D_{ij}, j \in C_{-i}\} - D_{iC_i}}{\max(\min\{D_{ij}, j \in C_{-i}\}, D_{iC_i})} \quad (4)$$

SC will be a value falling into the range of −1 to 1. If SC is close to 1, it means a state element is assigned to an appropriate cluster and is "well clustered." If SC is about 0, it means that the element could be assigned to another closest cluster as well, and the element lies equally far away from both clusters. If SC is close to −1, it means that the element is "misclassified" and is merely somewhere in between the clusters.

More recently, the so-called gap statistic has been proposed in the statistical literature as a measure for determining the number of clusters (Everitt et al., 2011). Tibshirani and his colleagues (2001) developed an approach that formalized the idea of finding an "elbow" in the plot of the optimized cluster criterion

against the number of clusters k. The basic idea of this approach is to standardize the graph of $\log(W_k)$ against the number of clusters, where W_k (an overall average within the cluster sum of squares) is a cluster criterion that has been minimized for k clusters by comparing it with its expectation under an appropriate null reference distribution. For this purpose, letting E_n^* denote the expectation under a sample size of n from the reference distribution, the optimal value for the number of clusters is then the value k for which the "gap" is the largest:

$$Gap_n(k) = E_n^*\{\log(W_k)\} - \log(W_k) \quad (5)$$

In Eq. 5, k is the number of clusters, n is sample size, and W_k denotes an overall average within the cluster sum of squares. Those interested in the theoretical details of this method can refer to the original paper (Tibshirani et al., 2001).

Cluster Algorithms

Generally, clustering algorithms can be categorized into two groups: hierarchical clustering and nonhierarchical clustering. In hierarchical clustering, there is no need to preset the number of clusters. The K-means algorithm, one of the most popular forms of nonhierarchical clustering, on the other hand, needs to specify the number of clusters arbitrarily. However, it is a faster and more reliable method, especially for applications with large, high-dimensional data sets. Furthermore, the K-means algorithm repeatedly reassigns elements to clusters so the same element can move from cluster to cluster during the analysis. While in hierarchical clustering, elements are added only to existing clusters and are forever captive in their cluster, only with a widening circle of neighbors. Considering the advantages and disadvantages of the hierarchical and nonhierarchical clustering, the proposed methodology for this study is to define the number of clusters by first using the silhouette and gap-statistic measures, and then using the K-means procedure to actually form the clusters.

A major limitation of existing clustering used in TOD breakpoints identification is that the merging schemes fail to take into account special characteristics of individual clusters, which cause outliers to occur when underlying features of the data, such as the time feature, are ignored. Thus, an enhancement of cluster analysis for identifying optimal breakpoints is to take the time of traffic occurring as one of the dimensions and include it in the clustering.
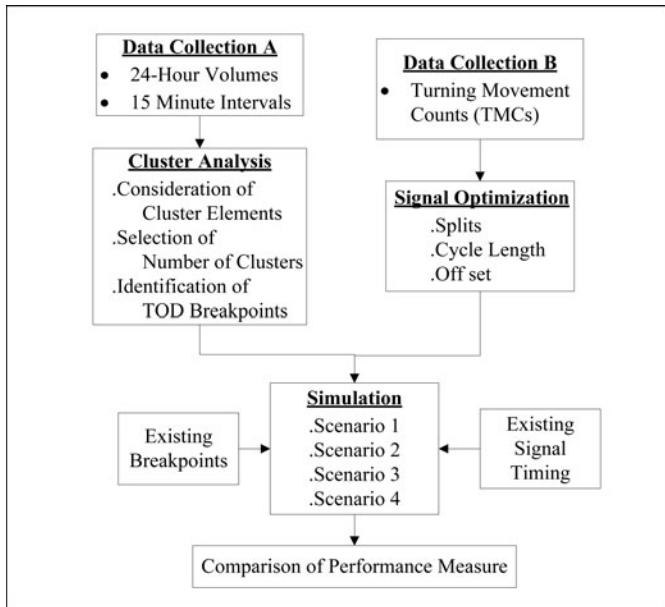
Figure 1   Proposed procedure for determining TOD breakpoints plans.

Proposed Procedure

Based on the concepts and considerations already described, an effective procedure for the development of the TOD signal timing plans formed in this research is shown in Figure 1. Each of the key components of the proposed procedure is described in detail as following.

Data Collection

Two kinds of traffic volume data involved in this study are shown in Figure 1. The first is 24-hour volumes, which were collected every 15 min in representative segments throughout the entire corridor. The 24-hour volumes are used in cluster analysis to determine TOD breakpoints. Another kind of data is turning movement counts (TMCs), which can be collected for 24 hours but are generally collected for 6 to 8 hours in different 2- to 3-hour time blocks of a day to reduce costs. TMCs are important inputs in both signal timing optimization and performance evaluation with simulation software. Besides the traffic volume data, information related to current breakpoints and existing signal timings was collected and used to create the scenario in simulation for demonstrating the performance of different breakpoints plans.

Enhanced Cluster Analysis

Based on the 24-hour volume data, enhanced cluster analysis was conducted with the following steps. First, proper cluster elements were selected. With consideration of the coordination of a corridor, bidirectional 24-hour volume samples from representative segments were selected as cluster elements. More importantly, a time variable was introduced as one dimension to remove certain outliers in the cluster analysis. Next, the silhouette and gap-statistic measures were used to find the appropriate number of clusters. Based on the results of the previous step, the K-means procedure was adopted to actually form the clusters. Given a specific threshold, all elements were assigned to the nearest cluster seed and new seeds were computed. Elements were then reassigned in successive steps, when necessary.

Signal Timing Optimization

Synchro was used to determine the macroscopic level of service (LOS) and delays of intersections along the corridor during the signal timing optimization process. In the context of signal timing plans, the following three parameters were of particular importance and were the focus of our optimization efforts: cycle length—the time required for a complete sequence of signal indications and signals in an actuated coordinated system, which should all operate under the same background cycle length; splits—the time assigned to a phase (green and the greater of the yellow plus all-red or the pedestrian walk and clearance times) during coordinated operations, which may be expressed in seconds or percentages; and offset—the time relationship between coordinated phases at subsequent traffic signals (Koonce et al., 2008). The optimization of these parameters was performed by using standard optimization techniques. After the signal timings were developed for each of the TOD intervals, the timing parameters were then entered into an Excel spreadsheet for the simulation and validation step.

Simulation and Validation

A microscopic traffic simulation tool was used to simulate and identify issues that may not have been fully realized with a macro-level model. Simulation, using microsimulation software such as CORSIM, requires three types of information: roadway geometry, traffic volume, and signal timing. Geometric information was collected from Google Earth and field notes. Data for 15-min TMCs from each time interval during selected periods supported the traffic volume requirement. A spreadsheet storing the signal timing parameters in the previous section was used to provide the third type of information. To validate this enhanced cluster analysis, four different scenarios were created and are described in the section of case study. The existing breakpoints and existing signal timing plans were both collected to perform as the baseline scenario. Finally, our method's effectiveness was evaluated by comparing its performance measures in the different scenarios.

CASE STUDY

Data Collection

Traffic data is the backbone of transportation analysis. To identify TOD breakpoints under various traffic conditions, continuous, stable, and accurate traffic samples at study sites are
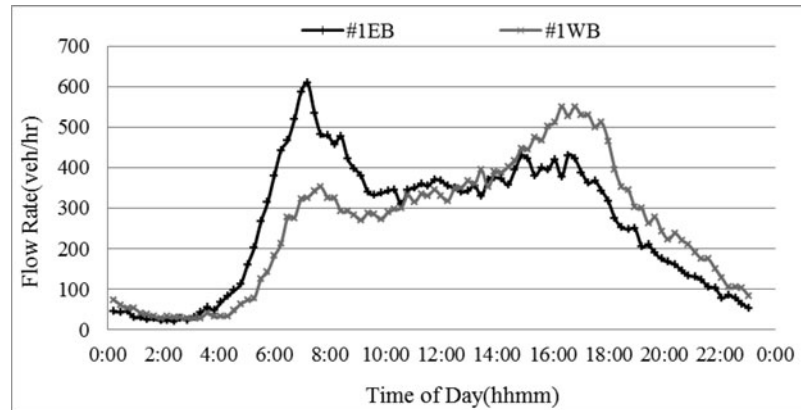
Figure 2   Weekday average hourly volumes for segment 1.

needed to cover all traffic conditions in a sequential fashion. Loop detectors are the most commonly used detection technologies for traffic data collection. However, the installation and replacement of loop detectors can be very disruptive to traffic. It is also extremely difficult to determine turning movement counts from loop data at intersections. Historically, turning movement counts were obtained manually by having observers at various spots of the intersections. In this study, traffic data were collected from nonintrusive data collection platforms, which offers a significant reduction in risk exposure over traditional segment count methods as well as a moderate reduction in risk exposure for traditional turning movement count methods.

24-Hour Volumes (Segments Counts)

Given the reduced risk exposure to personnel, the flexible installation options, and the high degree of accuracy, Wavetronix Smartsensor HD units were used to collect 24-hour approach counts in this study. Wavetronix Smartsensor HD units are portable, nonintrusive, and use dual-radar technology to detect traffic. This platform has a patented autoconfiguration process that defines the roadway cross-section and direction of vehicles in each lane. Traffic counts were collected from March 9, 2010, to March 15, 2010, at nine segments along Hillsborough Avenue, which is a major east–west arterial located in Hillsborough County, Florida. The corridor, which is approximately 15 miles in length, is primarily a six-lane divided arterial. Figure 2 shows an example of weekday average hourly volumes in one segment for both directions. It is observed that AM peak occurs eastbound while PM peak occurs westbound, which demonstrates the directional traffic on that corridor. Thus, we conduct cluster analysis separately for westbound and eastbound traffic.

Turning Movement Counts (TMCs)

TMCs at intersections were obtained with Miovision Video Collection Units (VCUs), which use digital video recordings to capture all vehicle turning movements. TMCs were recorded and stored on a Secure Digital (SD) card and then uploaded via office computer to the Miovision Web server (bandwidth dependent). The additional benefits of easy deployment, reduced requirements for trained staff, and the unique aspect of an audit trail make Miovision a viable alternative to traditional manual turning movement counting. In this study, TMCs were collected in the same corridor from March 23, 2010, to March 30, 2010.

Existing Breakpoints and Traffic Signal Timings

Existing data of breakpoints and traffic signal timings were obtained from the ATMS.now server of Hillsborough County and from City of Tampa traffic engineering personnel. Only weekday (Monday–Friday) signal timing plans and traffic conditions are compared in this study.

Results and Discussions

Consideration of Cluster Elements

As stated in the Methodology section, taking the time variable into account allows advanced clustering to operate successfully on temporal data sets that are available in metric spaces. Thus, time variables numbered from 1 to 96, corresponding with 96 sets of 15-min volume data over a 24-hour period, were considered as one dimension of inputs in this advanced cluster analysis. To account for the differences in scale between volumes and the time variable, the cluster elements were standardized by using Eq. 1 defined earlier. By following the procedure proposed earlier, the cluster membership value for each state exported from SAS outputs was plotted against TOD in Excel, as exemplified in Figure 3, to determine the TOD intervals.

Figure 3 shows a sample clustering outcome with the time of traffic occurring as one dimension (b) versus without (a). The clustering with the time consideration demonstrates a refined TOD timing plan scheme and the outliers are significantly reduced. In the particular case shown in Figure 3, the number of transitions with the consideration of time of traffic occurring is six, while it is nine without consideration of the time. Adding the dimension of the time of traffic occurring refines the
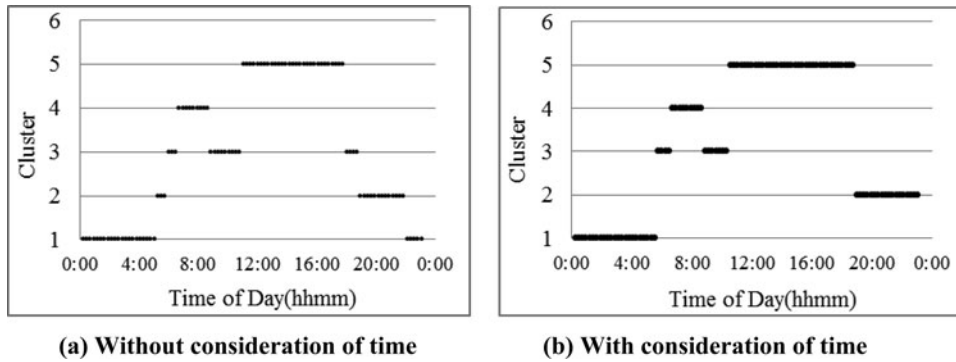
(a) Without consideration of time   **(b) With consideration of time**

Figure 3   TOD breakpoints with or without consideration of time.

cluster analysis and leads to a lower transition cost, besides other performance improvements that we elaborate later.

Selection of Cluster Numbers

To determine the optimal number of clusters, the silhouette coefficient and gap values were calculated by using MATLAB and R codes, respectively. Eastbound and westbound traffic samples are considered separately because of the directional characteristics of local traffic. Figure 4 shows the silhouette coefficient and the gap values against the number of clusters in our case study.

As shown in Figure 4, both the largest silhouette coefficient and gap values occur when the number of clusters is four. This is indeed consistent with existing knowledge on daily traffic where free-flow traffic, peak-hour traffic (congestion), and mid-

day traffic (transition) are among the most commonly observed traffic phenomena.

Identification of TOD Breakpoints

The K-means clustering successfully identifies the traffic patterns based on the average weekday 15-minute traffic volumes and the time that traffic is occurring, as displayed in Figure 5. Four clusters, as determined in the previous subsection, were conducted to find appropriate representation of natural volume groupings over a 24-hour period. Clusters 1 and 2 both represent the free (uncoordinated) operation of each intersection in the corridor over the night time, cluster 3 represents the mid-day period, and cluster 4 represents the a.m./p.m. peak periods.

It is assumed that the timing plans associated with combining eastbound and westbound breakpoints are subject to analysis and
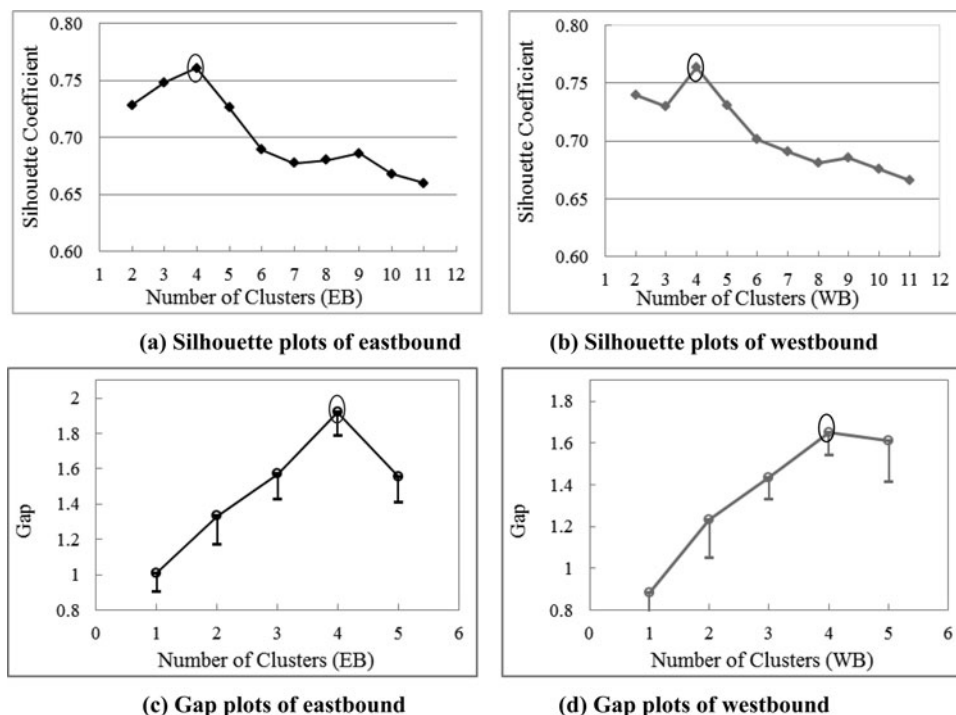

(a) Silhouette plots of eastbound   **(b) Silhouette plots of westbound**

(c) Gap plots of eastbound   **(d) Gap plots of westbound**

Figure 4   Silhouette coefficient and gap as a function of number of clusters.

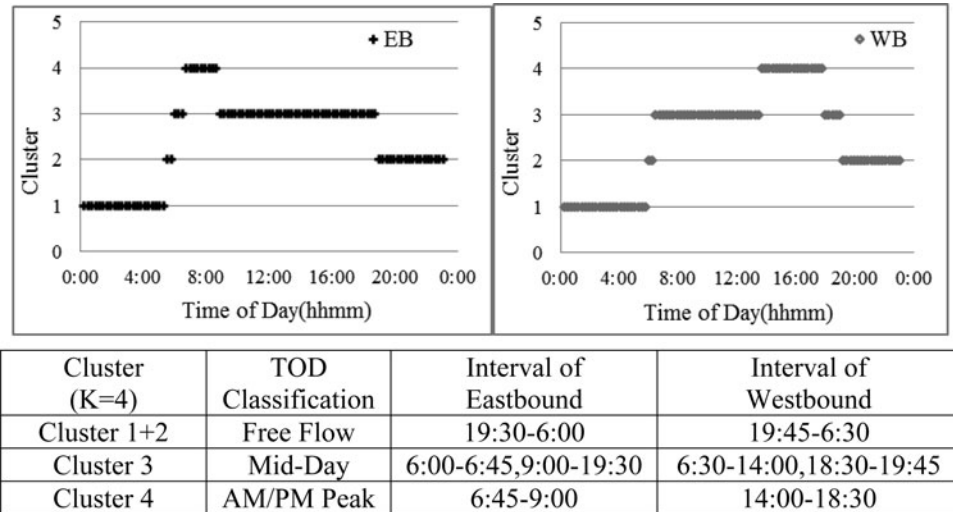| Cluster (K=4)	TOD Classification	Interval of Eastbound	Interval of Westbound
Cluster 1+2	Free Flow	19:30-6:00	19:45-6:30
Cluster 3	Mid-Day	6:00-6:45,9:00-19:30	6:30-14:00,18:30-19:45
Cluster 4	AM/PM Peak	6:45-9:00	14:00-18:30

Figure 5 TOD breakpoints and TOD interval classifications.

optimization. By combining the results of the enhanced cluster analysis for both directions shown in Figure 5, TOD intervals were developed over a 24-hour period for simulation and validation. The outcomes are compared with existing breakpoints and TOD intervals, which were obtained from the ATMS.now server. The existing TOD breakpoints were determined by traffic engineers in a traditional way, that is, plotting aggregate traffic volumes over the course of a day and then identifying the points of the changes in traffic volume by observing the plots.

Simulation and Improvement

It is important that the clusters obtained from the K-means clustering be validated. Thus, it is necessary to develop a traffic operation model of the corridor to evaluate the overall performance of our new method by comparing it to the performance of existing signal plans. Four scenarios were generated for the simulation analysis to fully investigate the effectiveness of the improved method. The first scenario represents a baseline situation where simulations were performed according to the existing breakpoints and existing signal timings. The second scenario kept existing signal timings, but adopted new breakpoints obtained from applying the new method. The third scenario kept existing intervals, but used new signal timings from the Syn-

chro optimization. These two hybrid scenarios were conducted to compare measures of performance with the baseline scenario by adopting new TOD breakpoints and new signal timings, respectively. The last scenario resulted from the application of the proposed procedure with both TOD intervals being identified from the advanced cluster analysis and the optimized signal timings. CORSIM simulations for different scenarios were conducted in multiple time periods with 15-min durations in each period. Figure 6 shows a snapshot of the corridor with six intersections in our case study that was simulated in CORSIM.

Due to the stochastic nature of traffic flow, it is necessary to run CORSIM multiple times while varying the random number seeds to gain an accurate reflection of the test case's performance. In this study, each TOD scenario was simulated with 30 runs. The average number of vehicles completing trips in each scenario for the 30 runs was 14,194. Additionally, the simulation modeling shows that all v/c ratios are less than 1. This means that the actual volumes did not reach capacity in our case study. During this study's multiple CORSIM runs, transitions between TODs were explicitly modeled using a short-way transition method, which was approved to be the best transition method according to Yun et al. (2008). The short-way method in CORSIM is a composite of the "Add" and "Subtract" methods. At the controller's yield point in each cycle during the transition,
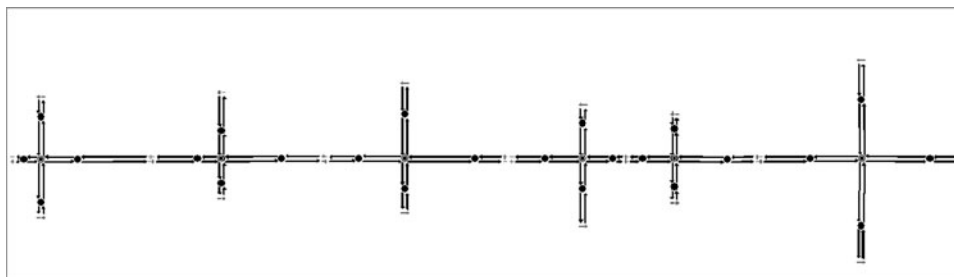

Figure 6 Hillsborough Avenue screen capture of simulation.

Table 3 Comparisons of performance measures.

| Measures of performance | Scenario 1 | | Scenario 2 | | Scenario 3 | | Scenario 4 | |
| | Existing TOD breakpoints and existing signal timing plans | | New TOD breakpoints and existing signal timing plan | | Existing TOD breakpoints and new signal timing plan | | New TOD breakpoints and new signal timing plan | |
	Mean	STD	Mean	STD	Mean	STD	Mean	STD
Average speed (mph)	21.63	0.275	22.32	0.320	22.73	0.271	23.28	0.320
Move/total (%)	52.58	0.646	54.26	0.759	55.26	0.640	56.59	0.758
Delay (min/mile)	1.32	0.035	1.23	0.038	1.18	0.031	1.12	0.035
Delay (sec/vehicle)	19.50	0.511	18.22	0.564	17.51	0.461	16.58	0.520

Note. CORSIM divides total travel time into two additive components: move time and delay time. Move time is the theoretical time it would take to traverse a link at the free-flow speed. The total travel time is the amount of time it actually takes to traverse a link. The delay time is then the difference in the observed total travel time and the theoretical move time.

Table 4 Paired samples test.

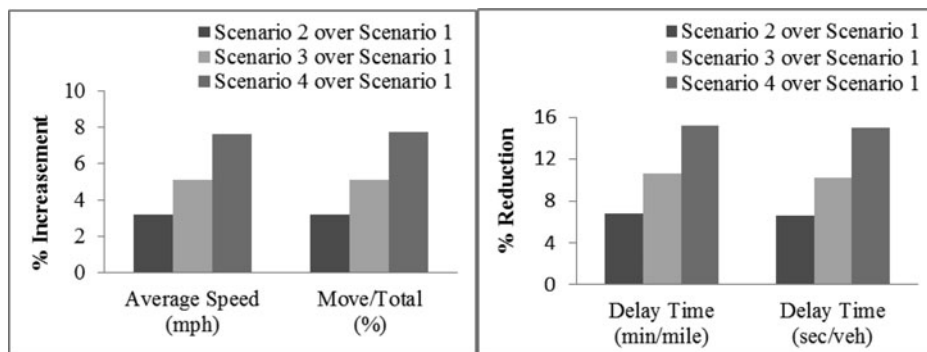| Pair of scenarios | Paired differences | | | | | | | |
| | Mean | SD | SE | 95% Confidence interval of the difference | | t | df | Significance (two-tailed) |
				Lower	Upper			
Average speed								
1 and 2	−.6938	.1293	.0236	−.7421	−.6455	−29.381	29	.000
1 and 3	−1.1017	.1572	.0287	−1.1604	−1.0430	−38.378	29	.000
1 and 4	−1.6516	.1751	.0320	−1.7170	−1.5862	−51.657	29	.000
Move/total								
1 and 2	−1.6765	.3167	.0578	−1.7948	−1.5582	−28.993	29	.000
1 and 3	−2.6771	.3792	.0692	−2.8187	−2.5355	−38.670	29	.000
1 and 4	−4.0010	.4236	.0773	−4.1680	−3.8517	−51.846	29	.000
Delay (min/mile)								
1 and 2	.0858	.0155	.0028	.0800	.0916	30.242	29	.000
1 and 3	.1345	.0194	.0035	.1272	.1417	38.017	29	.000
1 and 4	.1966	.0203	.0037	.1890	.2042	53.140	29	.000
Delay (sec/vehicle)								
1 and 2	1.2726	.2306	.0421	1.1865	1.3588	30.227	29	.000
1 and 3	1.9902	.2859	.0522	1.8835	2.0970	38.131	29	.000
1 and 4	2.9117	.2975	.0543	2.8006	3.0227	53.607	29	.000

Figure 7 Percentage changes of new breakpoints and signals over existing ones.

the transition manager computes the local zero difference. The transition manager then determines whether it is shorter (less time) to make up this difference by implementing an "Add transition" or "Subtract transition" for the next cycle (MacTrans, 2008).

The comparison of mean performance measures of four scenarios is given in Table 3. Additionally, the paired samples *t*-test was conducted to test if there was a significant difference between the mean performance measures. According to the results of the paired samples *t*-test, the mean performance measures under different scenarios are statistically significantly different. The higher average speed in the second scenario (compared with the baseline of Scenario 1 demonstrates that in this case study, the new TOD breakpoints obtained from the proposed method lead to better performance of the corridor. The comparison also demonstrates the benefit of refining signal time plans, given new TOD breakpoints. The combined effect leads to a nearly 8% increase in average traffic speeds throughout the corridor. The percentage changes of performance measures in different scenarios are depicted in Figure 7, with Scenario 1 as the base. As shown in Figure 7, the significant increase in average speeds and the move/total ratio (the ratio of the theoretical move time to the actual travel time for vehicles in the network) indicates that the developed signal system effectively services more vehicles through the corridor. Furthermore, delay times in minutes per mile and in seconds per vehicle are both significantly reduced in the Scenario 4.

CONCLUSION AND FUTURE WORK

This study developed a cluster analysis-based procedure that identifies TOD breakpoints for coordinated semiactuated traffic signal systems using continuous traffic data obtained through innovative, nonintrusive collection techniques. A novel modification, which proposes that time of traffic occurring be taken into account as a dimension, addresses the shortcomings of previous clustering approaches. The signal timing plans for the recommended TOD intervals were developed and evaluated in the simulation analysis. The results of a case study for a corridor located in Tampa, FL, demonstrated that the proposed method significantly improved the performance of the corridor.

A further step could be to develop several timing plans for the recommended TOD intervals, while considering the operational tradeoff between directional traffic flows. Next, their performance should be evaluated so that the most effective timing plans can be developed. Furthermore, efforts are needed to demonstrate the best way to estimate the necessary number of clusters while simultaneously considering both traffic flow directions. A sensitivity analysis should be performed for a variety of cluster numbers for this purpose. Also, for corridors with a large number of intersections, the dimension of the data set used for cluster analysis will be considerably higher. Due to the inherent difficulties encountered when working with high-dimensionaldata, innovative methods that can be used to convert multidimensional variables into one scalar are worth exploring.

REFERENCES

Abbas, M. M., & Sharma, A. (2005). Optimization of time of day plan scheduling using a multi-objective evolutionary algorithm. *Proceedings of 84th Annual Conference of Transportation Research Board of the National Academies*, January 2005, Washington, DC.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd edition). New York, NY: John Wiley& Sons.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster analysis* (5th ed.). London, UK: John Wiley & Sons, Ltd.

Holm, P., Tomich, D., Sloboden, J., & Lowrance, C. (2007). *Traffic analysis toolbox volume iv: Guidelines for applying corsim microsimulation modeling software* (No. FHWA-HOP-07-079). Retrieved from http://ops.fhwa.dot.gov/trafficanalysistools/tat_vol4/vol4_guidelines.pdf

Hauser, T. A., & Scherer, W. T. (2001). Data mining tools for real-time traffic signal decision support & maintenance. In *Systems, Man, and Cybernetics, 2001 IEEE International Conference* (Vol. 3, pp. 1471–1477), Tucson, AZ.

Koonce, P., Rodegerdts, L., Lee, K., Quayle, S., Beaird, S., Braud, C., . . ., & Urbanik, T. (2008). *Traffic signal timing manual* (No. FHWA-HOP-08-024). Retrieved from ops.fhwa.dot.gov/publications/fhwahop08024/fhwa_hop_08_024.pdf

Lee, J., Kim, J., & Park, B. B. (2011). A genetic algorithm-based procedure for determining optimal time-of-day break points for coordinated actuated traffic signal systems. *KSCE Journal of Civil Engineering*, **15**(1), 197–203.

MacTrans. (2008). *CORSIM reference manual: Version 6.1*. Gainesville, FL: University of Florida.

Park, B., Lee, D. H., & Yun, I. (2003, October). Enhancement of time of day based traffic signal control. In *Systems, Man and Cybernetics, 2003 IEEE International Conference* (Vol. 4, pp. 3619–3624), Washington, DC.

Park, B. B., & Lee, J. (2008). A procedure for determining time-of-day break points for coordinated actuated traffic signal systems. *KSCE Journal of Civil Engineering*, **12**(1), 37–44.

Ratrout, N. T. (2010). Subtractive clustering-based K-means technique for determining optimum time-of-day breakpoints. *Journal of Computing in Civil Engineering*, **25**(5), 380–387.

Ratrout, N. T. (2011). Developing optimal timing plans for cyclic traffic along arterials using pre-timed controllers. *Urban Transport XVII: Urban Transport and the Environment in the 21st Century*, **116**, 367.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Smith, B. L., & Venkatanarayana, R. (2005). Realizing the promise of intelligent transportation systems (ITS) data archives. *Journal of Intelligent Transportation Systems*, **9**(4), 175–185.

Smith, B. L., Scherer, W. T., Hauser, T. A., & Park, B. B. (2002). Data-driven methodology for signal timing plan development: A computational approach. *Computer−Aided Civil and Infrastructure Engineering*, **17**(6), 387–395.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**(2), 411–423.

Wang, X., Cottrell, W., & Mu, S. (2005, September). Using k-means clustering to identify time-of-day break points for traffic signal timing plans. In *Intelligent Transportation Systems, Proceedings, 2005 IEEE* (pp. 586–591), Vienna, Austria.

Xia, J., & Chen, M. (2007). Defining traffic flow phases using intelligent transportation systems-generated data. *Journal of Intelligent Transportation Systems*, **11**(1), 15–24.

Yang, L. C., Jia, L., Kong, Q. J., & Zhu, W. X. (2006). Method of automatic programming traffic intervals based on artificial immune algorithm. *Kongzhi Lilun yu Yingyong/Control Theory & Applications*, **23**(2), 193–198.

Yun, I., Best, M., & Park, B. B. (2008). Traffic controller performance during time-of-day transition at coordinated actuated signal systems. *Transportation Research Record: Journal of the Transportation Research Board*, **2080**, 92–99.