Final report of ITS Center project: Signal system data mining.

UVA Center for Transportation Studies

A Research Project Report

For the Center for ITS Implementation Research

A U.S. DOT University Transportation Center

Signal System Data Mining

Authors

Trisha Hauser
Dr. William T. Scherer
Dr. Brian L. Smith

Center for Transportation Studies
University of Virginia
Thornton Hall
351 McCormick Road, P.O. Box 400742
Charlottesville, VA 22904-4742
804.924.6362

September 2000
UVA-CE-ITS_01-3

Disclaimer

ABSTRACT

Intelligent transportation systems (ITS) include large numbers of traffic sensors that collect enormous quantities of data. The data provided by ITS is necessary for advanced forms of control, however basic forms of control, primarily time-of-day (TOD) which are prevalent in the United States do not directly rely on the data. Thus sensor data is typically unused and discarded in this country. The sensor data is in fact capable of providing abundant amounts of information that can aid in the development of improved TOD signal timing plans. Data mining tools are necessary to extract the information necessary from the data to improve on timing plan development and in turn would allow the timing plan development and monitoring process to be automated. This paper describes a research program that is investigating the application data mining tools, including statistical clustering and classification techniques to aid in the development of traffic signal timing plans. Specifically, a case study was conducted that illustrated that the use of Hierarchical Cluster analysis can be used to identify temporal interval break points that support the design of a time-of-day (TOD) signal control system. The cluster analysis approach was able to utilize a high-resolution system state definition that takes full advantage of the extensive set of sensors deployed in a traffic signal system. Finally, the case study also demonstrated that a Classification and Regression Tree (CART) could be developed that can be used to automatically monitor the quality of TOD intervals as traffic conditions change through time. The results of this research indicate that advanced data mining techniques hold high potential to provide automated techniques that assist traffic engineers in signal control system design and operations.

INTRODUCTION

It has been argued that traffic signal systems represent the first widespread deployment of intelligent transportation systems (ITS). Modern signal control systems are highly complex, relying on sensors, advanced communications networks, and sophisticated firmware and software. Advanced forms of signal control, such as second and third generation control, are dependant on the sensor data supplied by ITS. However basic forms of control such as time-of-day (TOD) do not rely on the sensor data for operation. These basic forms of control are in fact the most widely used methods of traffic signal control in this country due to limited funding for the Department of Transportation and the difficulty in maintaining the sensors for support of advanced control. These signal control systems are collecting enormous quantities of traffic flow data in an attempt to provide information for the support and improvement of signal timing operations. Unfortunately, due to limited storage resources, the lack of available analysis tools and the fact that the sensor data is not necessary for the support of TOD signal control, the vast majority of signal control systems in the United States do not archive this data for an appreciable period of time. It is in fact plausible to utilize the sensor data not only for advanced forms of control, but also for the most common method of signal control in this country, being TOD. Thus, there is a need to develop analysis tools that demonstrate the value of this data, and justify the design of systems with increased storage capabilities.

Tools used to analyze and extract *information* from large sets of *data* are generally classified as "data mining" tools. This paper describes an on-going research effort that is devising a procedure for developing, implementing and monitoring traffic signal timing plans using available data mining tools. The underlying premise of the research is that the data collected by signal control systems can be used to improve system design and operations for the current methods of use. The data mining tools that serve as the foundation for the proposed procedure for signal plan developments are Hierarchical Cluster analysis and Classification and Regression Trees (CART). This paper describes this research by offering a background on signal timing plan development, considering system state definitions, and detailing a prototype application of Hierarchical Cluster analysis and CART at an intersection in Northern Virginia.

BACKGROUND – SIGNAL TIMING PLANS

The operation of a signal control system on an arterial corridor requires a timing plan for each signal in the corridor. A corridor timing plan consists of three main elements: cycle length, splits, and offsets *(1)*. The cycle length is the time required for one complete sequence of signal phases to rotate through the green time. The split refers to the percentage of a cycle length allocated to each of the various phases at an intersection in a signal cycle, where phase refers to the portion of a cycle allocated to any single combination of traffic movements simultaneously receiving the right-of-way *(1)*. Finally, the offset is the component of the signal-timing plan that coordinates a series of signalized intersections in a corridor or network. The offset is the time difference (in seconds or in percent of the cycle length) between the start of the green indication at one signal as related to the start of the green indication at the corresponding downstream signals *(1)*.

The most widely used method for timing plan selection and implementation in the United States is time-of-day, or TOD, where a pre-set plan is automatically used for a particular time interval *(1)*. TOD requires traffic engineers to develop signal timing plans that are effective for particular time intervals in a day. For example, an AM-peak plan that favors work-bound commuter traffic might be used from 06:00 – 09:30. The AM-peak plan would typically be developed using timing optimization tools such as SYNCHRO, based on vehicle hand-counts taken during a typical peak period at the intersection. Therefore, one will note that the challenge in designing a TOD system lies in identifying the appropriate intervals for plans, determining when a “typical” day occurs for volume counts and then developing effective plans to operate within each interval.

There exist a number of optimization tools to assist traffic engineers in developing timing plans for a particular set of operating conditions. However, few tools exist to help the engineer determine appropriate intervals, or to monitor an existing TOD system to ascertain if the conditions have changed sufficiently to require a new set of plans and/or intervals. The premise of this research is that using data mining tools such as statistical clustering and classification for timing plan development and implementation has high potential to address these needs.

TIME OF DAY INTERVAL IDENTIFICATION

The typical approach used to identify intervals for TOD systems is to plot aggregate traffic volumes over the course of a day, and then use judgement to identify significant changes in traffic volume that indicate a need for a different timing plan, and, therefore, an interval break point. It is important to note that the volumes used to identify time-of-day break points are bi-directional aggregate volume values. An example of this approach is illustrated in Figure 1, which depicts a daily aggregate volume plot at an intersection in Northern Virginia. The vertical lines in the graph show the times that the traffic engineers chose to transition between plans, the interval break points. Along this particular corridor, there exists an AM-peak plan that operates from 06:00 – 08:30, a Mid-day plan that operates from 08:30 – 15:00, a PM-peak plan that operates from 15:00 – 19:00, and an off-peak plan for the remainder of the day.

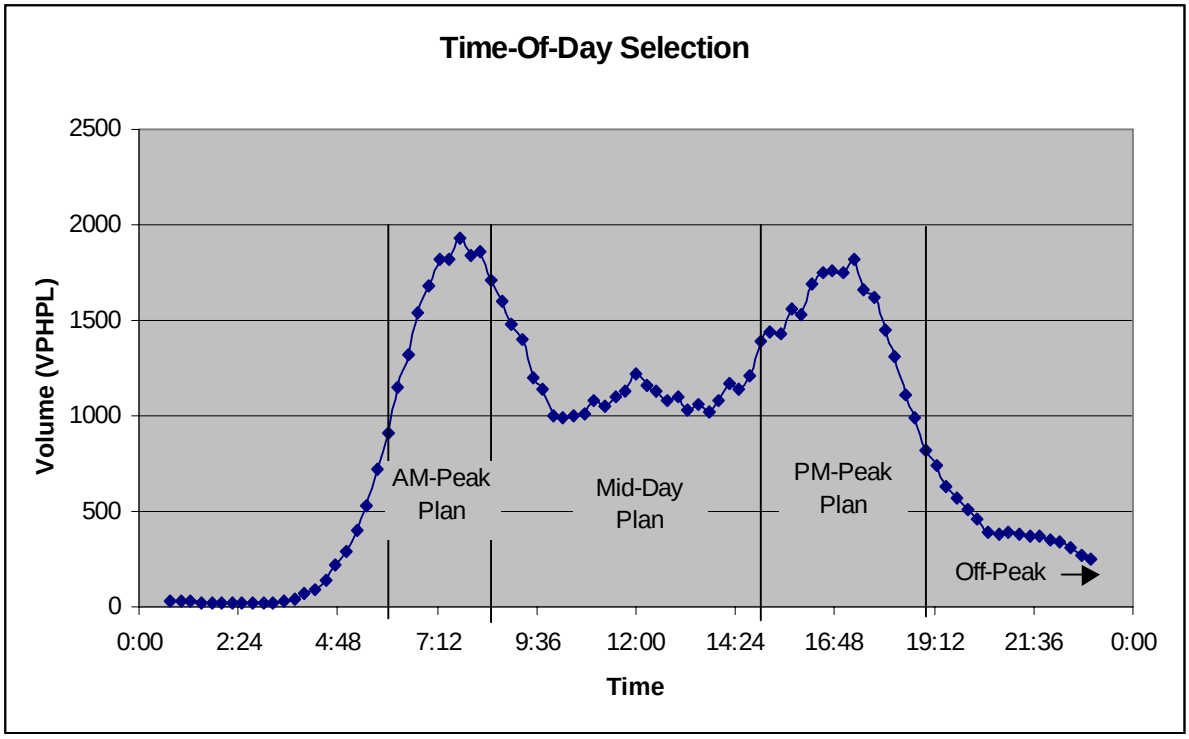


FIGURE 1 TOD Interval Identification

While this approach is intuitive, there are a number of areas of concern. First, the aggregation of only volume from traffic sensors (that typically measure volume, speed, and occupancy) in different directions (and, often, even lanes), to one aggregate volume measurement results in the loss of considerable information regarding the characteristics of the traffic conditions. In addition, as timing plans are developed for corridors, as opposed to single intersections, this loss of data resolution becomes more apparent. Finally, the visual selection of break points may be quite difficult for inexperienced engineers. These problems illustrate the need for automated data mining tools that are capable of identifying natural break points based on historical data, while using occupancies and/or speeds in addition to volumes for determination of break points, thus taking advantage of the large quantities of data collected by ITS.

STATE DEFINITION

TOD signal control is an example of a commonly used approach to system control known as state-based control. A "state" is an abstract representation of the condition of that system at some point in time. The defined state serves as a sufficient statistic for the condition of the system, *i.e.*, it contains all possible information regarding current status, propensity to change and response to external control, as well as the information necessary to evaluate the defined indices of performance for the system *(2)*. The concept of state-based control is to use a set of established rules or policies to guide the selection of a control strategy for a system as the system transitions from one state to another.

Clearly, the current practice of using aggregate volumes to define state, as described in the previous section, may be inadequate. Given that considerably more information is available to use in defining the state of the system, this research uses a more complete state definition based on the rawest form of data available from the system detectors to identify TOD intervals.

CLUSTER ANALYSIS

The concept behind TOD control is that traffic conditions during particular intervals of the day are roughly equivalent, and therefore a single timing plan can be used effectively throughout that interval. In other words, if traffic conditions are sampled at regular intervals, two samples, which will be referred to as cases, measured during the same TOD interval will be very similar. Cluster analysis is a statistical technique that has been developed to "group together" similar cases. Clustering algorithms are methods to divide a set of

n observations into *g* groups so that the members of the same groups are more alike than members of different groups or clusters *(3).* Thus, the premise of this research is that cluster analysis can be used to automatically group together similar samples of traffic conditions to identify TOD intervals.

With Hierarchical Cluster analysis, observed data points are grouped into clusters in a nested sequence of clusterings such that the algorithm starts with n clusters, each containing only one observation and joins the n clusters one at a time until only one cluster remains. The two closest clusters or observations are joined based on the measure of dissimilarity (d) chosen to be used, in the case the squared Euclidean distance.

$$\mathbf{d = \Sigma_{k=(1,9)}[(X_i^k - X_j^k)^2]}$$

The dissimilarity between each new cluster formed and any other observation or cluster is defined as the minimum distance between the two observations in the new cluster formed and any other observation or cluster. While the clusters are formed based on the minimized dissimilarity within clusters, the distance between clusters is maximized based on the squared Euclidean distance between cluster centroids. A minimum number of observations belonging to each final cluster formation is one constraint imposed on the cluster analysis such that clusters formed are valid based on a significant amount of observations, thus assuming clusters are not formed based on erroneous cases. This constraint also forces the time intervals formed by the cluster analysis to be of a significant duration, i.e., 30 minutes or greater. The time lost due to transition between timing plans is not accounted for in this research and should be further investigated in future research.

It is of importance with Hierarchical Clustering to determine the optimal number of clusters, for it is this number that represents the number of timing plans to develop based on the sensor data. In cluster analysis, the rules which determine the optimal number of clusters are called “stopping rules.” The cubic clustering criterion (ccc), a measure produced by the statistical software package, SAS, is the stopping rule implored in this research. The ccc is based on the R^2 value, where R^2 is the proportion of variance accounted for by the clusters, and it is based on the p value, where p is an estimate of dimensionality of the between cluster variation (5).

$$\mathbf{ccc = \{ln[(1 - E(R^2)) / (1 - R^2)\} * \{((np/2)^{.5}) / ((.001 + E(R^2))^{1.2})\}} \qquad (5)$$

The largest ccc value represents the most stable and meaningful level of the Hierarchical Cluster tree at which point the clusters are most representative of the timing plans and break points to be developed based on historical traffic conditions.

Cluster analysis is an ideal data-mining tool because the classes or groups that the data form are unknown, especially as the state definition is expanded to include an increasing number of variables. Cluster analysis uncovers these underlying patterns in the data and assigns each case to a group or cluster. Using the large set of historical data collected by signal systems, cluster analysis can be used to "mine" the data in order to help traffic engineers determine appropriate TOD intervals or break points to better develop signal timing plans.

RESEARCH CASE STUDY

The case study conducted in this effort utilized the resources of the University of Virginia's Smart Travel Laboratory. The laboratory is directly integrated with a number of traffic control systems operated by the Virginia Department of Transportation (VDOT). The data used in this effort was from VDOT's Northern Virginia signal system. This system uses inductive loop detectors positioned well upstream of intersections (system detectors) to collect basic demand data. Of the measures available from the detectors, volume and occupancy are the most commonly used variables in traffic control and the most important in traffic signal plan selection and development *(1).* Northern Virginia's inductive loop detectors aggregate volume and occupancy values every 15-minutes. It should be noted that while the system also collects average speed data, it does so by deriving speed from volume and occupancy values. Because the derivation of speed requires significant assumptions, the speed data has not proven reliable and is not used in this research.

The case study was conducted at a single intersection, Sunset Hills and Reston Parkway, the critical intersection in the Reston Parkway corridor in Northern Virginia. A corridor's critical intersection typically experiences the largest amount of traffic volume and has the most complex geometry. This is the case for the Sunset Hills and Reston Parkway intersection, which consists of 8 phases with multiple turning and through movement lanes. Figure 2 illustrates the intersection configuration, with the location of system detectors denoted by diamonds. A major shortcoming for this timing plan development procedure is that

system detectors are not available in all lanes, thus a representation of all movements form the historical data base may not be available at each intersection. A scaling factor is developed based on the existing hand-counted volumes to determine the ratios existing between through and turning movements and this scale factor is used to estimate movements where system detectors are absent based on the historical data available.

Due to the preliminary stage of this research, a small data set was used to validate the formation of the clusters. The volume and occupancy data were collected at 15-minute intervals from March 1, 2000 until June 26, 2000. From this data set, 124 distinct observations were selected representing all portions of a 24-hour period.

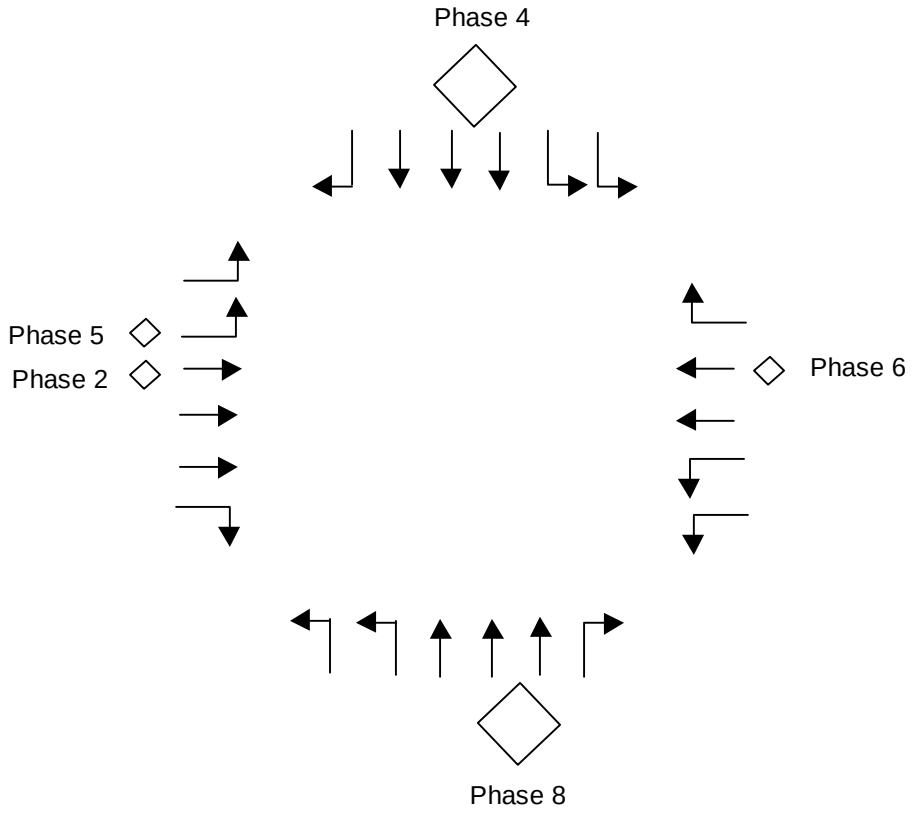


FIGURE 2 Sunset Hills and Reston Parkway intersection layout

State Definition

By considering the data collected by the system detectors in as high a resolution as possible, one can expect to better capture the nuances of the system's dynamic behavior. Therefore, the state definition used for this

case study is a vector of volume and occupancy measures for each directional phase movement at the Sunset Hills and Reston Parkway intersection. The directional phase movements are identified by their corresponding phase numbers, which are denoted in Figure 2. In addition, to account for the difference in scale between volume and occupancy measures, the values were standardized using a z score, which represents the number of standard deviations from the mean that each value lies. *Since volume and occupancy represent different traffic states, where occupancy values lie on a scale of 0 – 100 and volume values lie on a scale of 0 – 1900+, the standardization process is necessary to transfer these values to a uniform, meaningful scale. The different scales that represent the two variables reflect the differing nature of variables and the standardized value produced by each variable distinguishes the nature of the variable. This eliminates the need to assign weights to the state variables during the clustering procedure.* ***{"Not well explained?!"}*** For the scope of this research, the detectors were also weighted equally, however future considerations should include weighting cluster variables such as detectors and intersections to account for influence and importance of those factors in traffic flow through the corridor. Therefore, the state is as follows, with each variable number assigned according to its phase number:

$$X(t) = (V2, O2, V4, O4, V5, O5, V6, O6, V8, O8) \quad (2)$$

Where X(t) = system state at time t

V2 = standardized phase 2 volume at time t

O2 = standardized phase 2 occupancy at time t

V4 = standardized phase 4 volume at time t

O4 = standardized phase 4 occupancy at time t

V5 = standardized phase 5, left-hand-turn volume at time t

O5 = standardized phase 5, left-hand-turn occupancy at time t

V6 = standardized phase 6 volume at time t

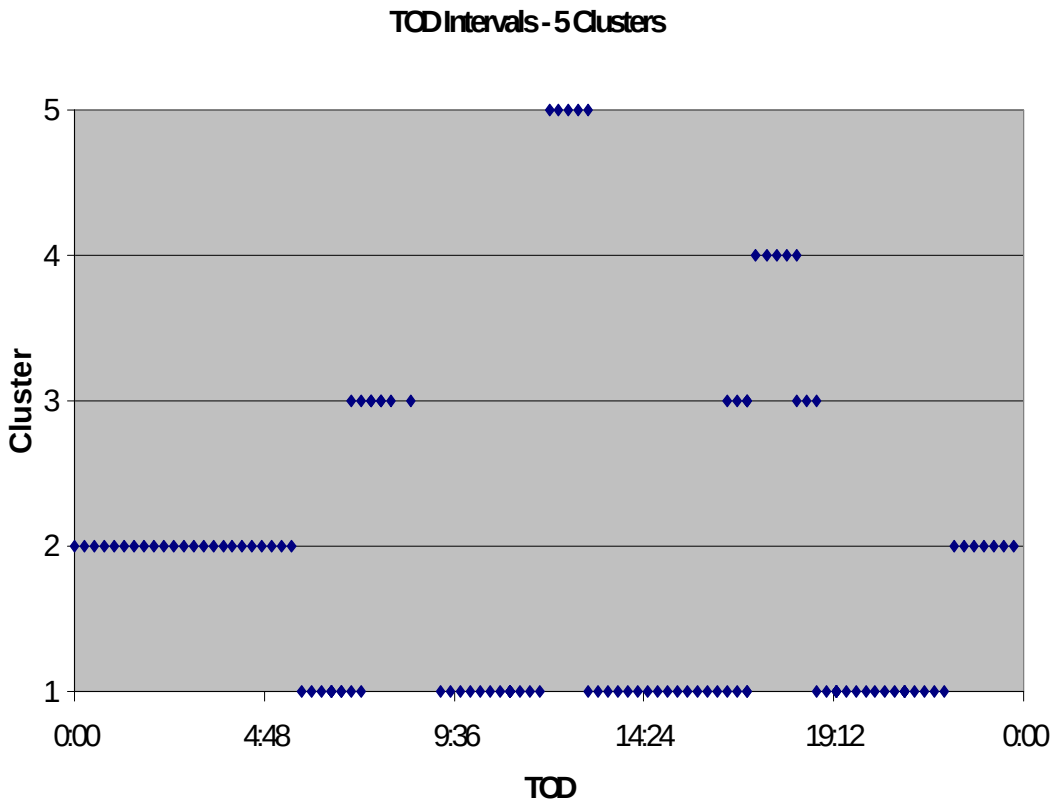
O6 = standardized phase 6 occupancy at time t

V8 = standardized phase 8 volume at time t

O8 = standardized phase 8 occupancy at time t

Cluster Analysis

VDOT is currently using four timing plans/intervals along the Reston corridor, an AM-peak plan, a mid-day-peak plan, a PM-peak plan and an off-peak plan. See Figure4. The Hierarchical Cluster analysis was performed on 124 observations taken from different days occurring during the 24-hour period of a day. The cubic clustering criterion (ccc) for this cluster analysis indicated that 5 clusters were optimal for this cluster tree formation, thus representing 5 timing plans to operate at that intersection. Based on the composition of each cluster, the ideal time-of-day for transition from one timing plan (associated with cluster *i*) to another timing plan (associated with cluster *j*) can be determined from the cluster analysis. Figure 3 illustrates which cluster (on the y-axis) each case falls within over the course of the day (x-axis). From figure 3, it is clear that the 5 timing plans represent uniquely occurring traffic conditions that arise during a 24-hour period. Cluster 1 represents the pre-AM, mid-day and pre-off peak periods of the day, cluster 2 represents the off peak period, cluster 3 represents the AM and pre/post PM peak periods, cluster 4 represents the PM peak period and cluster 5 represents the lunch peak period of the day. It appears that clustering on historical data illustrates a more refined TOD timing plan scheme. The different periods of the day that fall into similar clusters make sense in that traffic conditions would intuitively be similar at those times. For instance the periods occurring just before and after the major PM period would more match the conditions that exist during the AM period, thus these times being represented by a similar timing plan with refined TOD intervals. The cluster analysis tends to pick up on sensitivities such as these. The move to larger data sets may not produce as clean of a representation of TOD intervals as does figure 3. The idea of using such representations as Figure 3 for the use of determining TOD break points is to look for emerging trends produced by cluster analysis as represented by the majority of the historical data for timing plan development.



Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Pre-AM, Mid-Day, Pre-Off-Peak 5:45 – 7:00 9:00 – 12:00 13:00 – 16:30 18:45 – 22:00	Off Peak 22:00 – 5:30	AM, Pre/Post PM 7:00 – 9:00 16:30 – 17:00 18:15 – 18:45	PM Peak 17:00 – 18:15	Lunch Peak 12:00 – 13:00

FIGURE 3 TOD break points

Cluster analysis also identifies cluster centroid information, as seen in Figure 4. The volume data is necessary for the development of timing plans for each interval, and thus the values used to construct Figure 4 provide the appropriate data for use with timing optimization packages. These values should also provide a more accurate representation of actual traffic conditions for timing plan development versus using the traditional means of hand-counted volumes in traffic signal development. The centroid

representation of the clusters in Figure 4 also represent the uniqueness of the clusters formed and the mean volumes and occupancies occurring with each cluster. The error bars represent the standard deviation within the clusters.

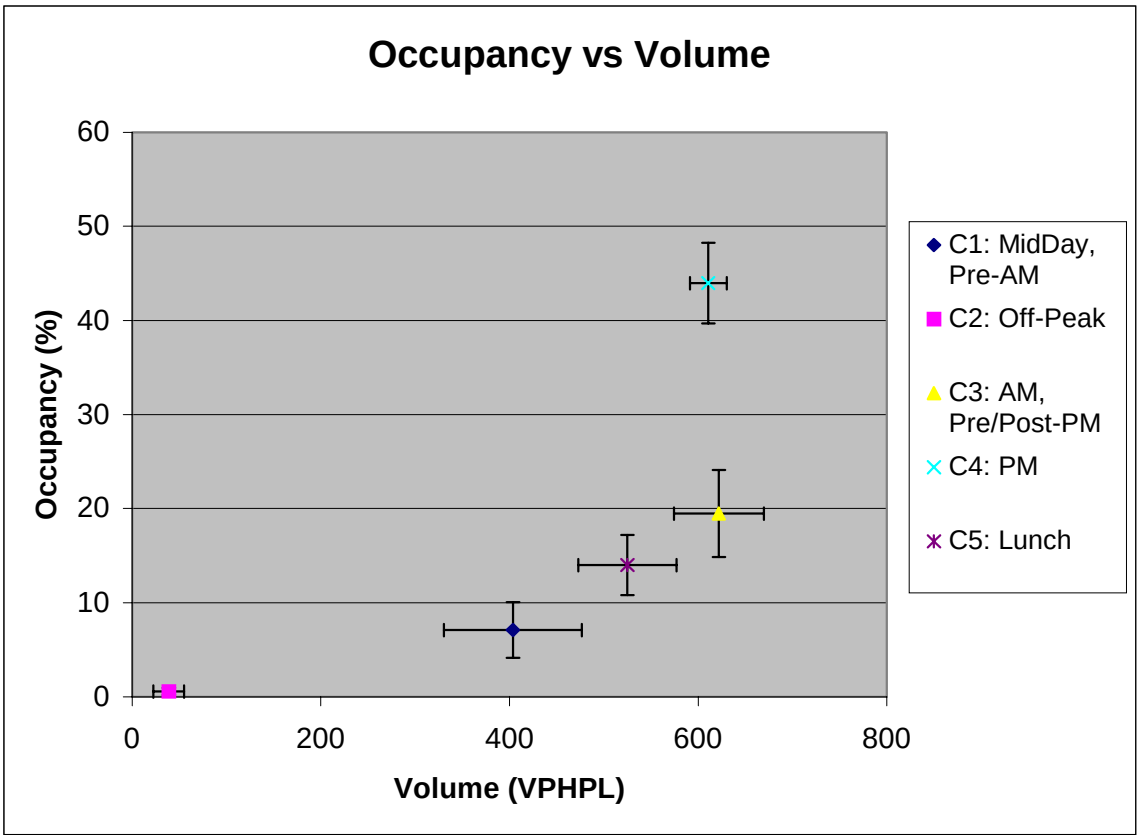


FIGURE 4 Cluster Centroids

Figure 5 depicts the comparison between TOD break points that are defined currently in Northern Virginia, and the TOD break points established by cluster analysis. A number of observations may be made when considering this figure.

- The transition to the AM-peak plan occurs at nearly the same time for both approaches. This indicates that a significant change in system state is clearly evident between 06:30 and 06:45.
- The cluster analysis approach clusters a large portion of the day into the Mid-day peak plan minus the transition to the lunch period. In addition, for a short period after the PM peak hours and before off

peak, around 18:45 – 22:00 and prior to the AM peak period, the system is again clustered into the Mid-day-peak plan. This may indicate that community traffic, for activities such as shopping and recreation, more closely matches Mid-day and pre-AM-peak traffic.

- The PM-peak period is relatively short, from roughly 17:00 - 18:15 representing the fine-tuning of the TOD break points according to traffic conditions.
- The shoulders occurring before and after the peak period are grouped with the AM peak plan suggesting that during these hours traffic more closely matches the patterns of AM peak traffic conditions and that the clusters are catching this "fine-tuning" for plans
- An additional plan was suggested for the lunch peak occurring from 12:00 – 13:00, suggesting that traffic conditions are unique from other periods as commuters travel to eating establishments during that hour

The interesting trend to note is that the cluster analysis resulted in intervals that make empirical sense based on general commuting trends. This illustrates the potential of this approach to automate aspects of TOD timing plan development. In addition, it is clear that the approach calls for slightly different transition times than using the traditional approach. This may be due to the fact that the cluster analysis combined with a higher resolution state was able to distinguish more subtle changes in system status. Finally, it should be noted that the cluster analysis approach calls for additional transition periods for the timing plans. Given the difficulty associated with transition in a coordinated system, this is a tradeoff that will be dealt with in future work, using simulation to compare the performance of these two designs.

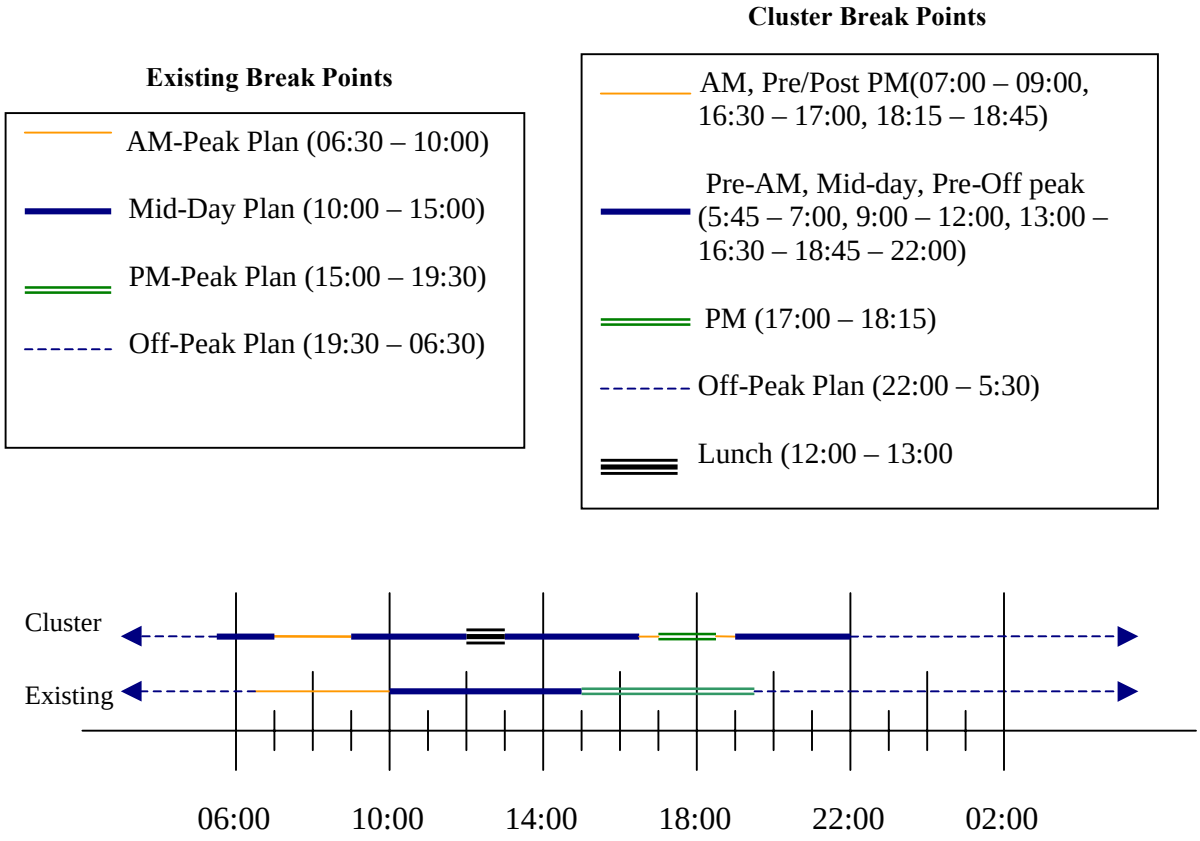


FIGURE 5 Cluster vs. existing interval break points

MONITORING OF TIMING PLAN EFFECTIVENESS

A significant challenge to effectively operating a TOD system is to ensure that the timing plans are refined to accommodate periodic changes in traffic conditions. Communities continually change, with new industrial, commercial, and residential areas being developed. This development leads to changes in traffic patterns, which in turn should lead to new plans for use in a TOD system. Once a set of plans has been designed, there is a need to automate the monitoring of these plans to ensure that they do not need to be changed. Using the results of the clustering approach in conjunction with a classification system can provide this capability.

Supervised classification is used to define rules for state classification into one of the 4 clusters identified using the Hierarchical Clustering approach. This allows one to check each new case of data

collected from the field to determine if it "classifies" in the cluster that is associated with the TOD plan in use. A series of misclassified states would warn the operator of the need to re-cluster for the development of new plans and interval break points. Misclassified states infer that traffic conditions have changed over time and the current plans may no longer be optimal.

Classification and Regression Trees (CART) is the predictive modeling statistical tool used to develop the classification rules in this research. CART is a robust data-analysis tool that automatically searches for important patterns and relationships in data sets and uncovers hidden structure *(3)*. CART uses a "no stopping" rule to determine the optimal tree. This method over-grows the tree and then prunes back to ensure that the tree does not stop growing prematurely *(3)*. CART implores binary splits on the data, which ultimately detect more structure without the risk of splitting the data too rapidly. CART also uses embedded tests such as cross-validation and test/training sets to ensure that the data is not overfit *(3)*. The selection of the optimal tree is performed by CART and not left up to the user, so the selection of the optimal classification rule is generated automatically. With the classification rule, each real-time volume and occupancy pair that occurs can be run through the rule and thus assigned a classification or cluster number that is appropriate for that particular traffic condition. If the traffic condition is assigned the cluster number for that plan which it is operating in, then the plan would be up-to-date. However, if an alternate plan than that which the current state is operating in is suggested by the classification rule, then it can be determined that the plan is not optimal and a series of these "misclassifications" would suggest the need for re-timing.

Using the data developed by the cluster analysis described in the previous section, the class variable for CART is defined as the cluster number. The classification model developed by CART used a test/training set to develop the model. CART used 67% of the data as the learn set to train the model and it uses the remaining 33% of the data to test the validity of the model. A preliminary tree was developed using CART, which created a classification rule that was 97.2% accurate on the test data set. This indicates that the use of CART allows for effective classification that may provide a powerful data-mining tool to continually check for deterioration in TOD timing plan effectiveness.

FUTURE WORK

Based on the promising results of the case study, this research will be extended. The first task to be conducted is the development of timing plans using optimization tools based on the cluster centroid volumes. Using these fine-tuned timing plans, simulation will be used to test the performance of the system using the TOD break points developed by cluster analysis for comparison with the existing TOD break points implemented along the Reston corridor. The performance of the developed clusters with the existing four timing plans implemented by VDOT will be investigated by comparing measures of performance such as total delay, number of stops, fuel emission and travel time.

The final stage of this research will be to perform clustering based on an entire corridor, not a single intersection as in the case study. The state definition used to form the clusters will be considerably larger, consisting of all available movements at all intersections in the corridor. The clusters will also be tested through simulation using only a select few intersections to discover the minimal number of input variables that still provide optimal results. The corridor to be analyzed in this study is the Reston Parkway corridor, which consists of 15 intersections and 82 system detectors.

CONCLUSIONS

ITS technology provides abundant amounts of data describing transportation system status. With such large databases, it becomes difficult to organize and keep track of the underlying meaning of the values in the database. Data mining tools are the keys to interpret such large amounts of data. With data-mining techniques such as clustering and classification, it is possible to improve signal-timing plans.

Here, it has been shown that clustering will find patterns in the data using a more detailed and descriptive state definition than the current practice. The clusters formed supply the time-of-day break points at which transition is optimal for the conditions that exist. Simulations run for preliminary purposes do in fact show that the current timing plans operating at the test intersection improve performance when run under the cluster times of day. Another data mining tool, classification is also useful for this process to classify future cases into specified clusters. This classification process alerts the traffic engineer of out-dated timing plans. The data mining classification and clustering procedures will greatly aid traffic

engineers in the development, selection and implementation of timing plans that will more efficiently control traffic.

REFERENCES

1. Case, French, Gordon, Haenel, Mohaddes, Reiss and Wolcott. Traffic Control Systems Handbook. In *Publication Number: FHWA-SA-95-032*, February 1996.

2. Puterman, M. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York, 1994.

3. Ripley, B. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, United Kingdom, 1999.

4. Mendenhall, W., Sincich, T. *A Second Course in Statistics: Regression analysis, 5th edition*. Prentice Hall, New Jersey, 1996.

5. Milligan, Cooper. *An Examination of Procedures for Determining the Number of Cluster in a Data Set*. Psychometrika, June 1985. Vol. 50, No. 2, pp. 159 – 179.