**Final report for ITS Center Project: Adaptive Signals Decision Support System.**

**UVA Center for Transportation Studies**

**A Research Project Report**

**For the Center for ITS Implementation Research**

**A U.S. DOT University Transportation Center**

# Data Mining Tools for the Support of Traffic Signal Timing Plan Development in Arterial Networks

Principal Investigators:

William T. Scherer,

Brian L. Smith

Graduate Assistant:
Trisha Ann Hauser

Center for Transportation Studies

University of Virginia

Thornton Hall

351 McCormick Road, P.O. Box 400742

Charlottesville, VA 22904-4742

804.924.6362

May 2001

UVA-CE-ITS-02-6

## 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 by providing historical data for automatic plan development and TOD interval identification. 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 rather than the time-consuming, intuition based practice currently implemented. This project describes research investigating the application of data mining tools, including statistical clustering techniques, to aid in the development of traffic signal timing plans. Specifically, a case study was conducted to illustrate that the use of hierarchical cluster analysis can be used to automatically identify temporal interval break points, based on the data, 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. Timing plans were developed based on the clustering results, providing enhanced TOD intervals and peak volumes, which were then tested through simulation and internal cluster validation, which proved that the use of data mining tools for plan development is beneficial. The results of this research indicate that advanced data mining techniques hold high potential

to provide automated techniques to assist traffic engineers in signal control system design, development and operations, the entire process of plan development that is currently practiced based on hand-counted volumes and single intersection TOD intervals.

# Table of Contents

# LIST OF FIGURES

# LIST OF TABLES

# Chapter 1. INTRODUCTION

## *1.1 Traffic Signal Systems and ITS*

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. 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 detector data for an appreciable period of time. This is unfortunate, especially since it is 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, TOD. Thus, there is a need to use analysis tools that demonstrate the value of this data, and justify the design of systems with increased storage capabilities.

### *1.2 Data Mining Tools*

Tools used to analyze and extract *information* from large sets of *data* are generally classified as "data mining" tools. This project describes research that is devising a procedure for developing, implementing and monitoring traffic signal timing plans using available data mining tools. The hypothesis 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 traffic control. The data-mining tool that serves as the foundation for the proposed procedure for signal plan developments is hierarchical cluster analysis. It will also be recommended that a second data-mining tool, classification, be used for monitoring plan effectiveness, however this project will not explore the use of classification in the maintenance of timing plans in depth. This project offers a background on signal timing plan development, with consideration of system state definitions, and detailing a proposed procedure for improved traffic control through the use of hierarchical cluster analysis with a case study at a corridor in northern Virginia. This case study shows that the sensor data provided by ITS holds valuable information regarding the behavior of traffic, capable of automatically generating TOD intervals for transitioning between timing plans as well as providing appropriate volume data for plan development during these automatically generated TOD intervals. The proposed procedure introduced in this project allows for automation of the entire signal timing plan process, which will save time for traffic engineers and improve travel conditions for commuters.

### *1.3 Existing Plan Procedures*

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 TOD 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 TOD intervals. Certainly, no tools exist to accomplish these tasks automatically. The premise of this research is that using statistical clustering and classification analyses in a data mining application has high potential to address these needs and allow for automated procedures, while utilizing the information stored in the data for optimal signal development and maintenance.

Clearly, the current practice of using single day, hand counted volumes to define the state for time-of-day (TOD) plan development may be inadequate. Given that considerably more information is available to use in defining the state of the system in electronic form, this research uses a more complete state definition based on a more refined form of data available from the system detectors to identify TOD intervals and develop more appropriate timing plans.

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 judgment in the identification of significant changes in traffic volume at the critical intersection that indicate a need for a different timing plan. It is important to note that the volumes used to identify TOD intervals are bi-directional aggregate volume values from the critical intersection. An example of this approach is illustrated in Figure 1, which depicts a daily aggregate volume plot at an intersection in Northern Virginia based on historical data.

The vertical lines in the graph show the times that the traffic engineers chose to transition between plans, the TOD intervals. These intervals rely heavily on the traffic conditions that exist at the critical intersection. The critical intersection is the signalized intersection in the corridor servicing the largest traffic demand. 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**

### *1.4 The Need for Improved Control*

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 TOD intervals may be quite difficult for inexperienced engineers, who ultimately spend much time developing and tweaking the plans and TOD intervals. These problems illustrate the need for automated data mining tools that take advantage of the large quantities of data collected by ITS. The use of cluster analysis addresses these issues and uses a more robust state definition based on historical data at all intersections and at all movements in the development of plans and TOD intervals. Figure 2 depicts the TOD intervals developed by a clustering procedure versus those developed manually as described above. It will become clear that the clustering TOD intervals are more robust based on the detector data rather than the one-day hand counts practiced currently. Sensitive traffic trends are detected by the clustering algorithm that occur over a 24-hour period and these TOD intervals developed via clustering will be investigated in detail in Chapter 5.

**Clustered TOD Intervals**

**Existing TOD Intervals**

**Figure 2. Existing TOD Intervals vs. Cluster Intervals**

It is also becoming increasingly vital to provide efficient and up-to-date signal control due to the decreasing availability of land for road expansion (6). With the explosion of population, industry and "suburbia," traffic conditions are becoming increasingly congested in many spreading areas and with this growth comes the need for more roads. However, the land is becoming less available and the zoning laws stricter, making it extremely difficult to build new roads. The cost of building new roads is also extremely expensive and traffic engineers are relied upon heavily to provide efficient

forms of traffic control to deal with growing traffic problems where new roads may be highly needed but nearly impossible to construct. Thus, the TOD signal procedures that have not changed much over the past decades need to be improved, with a more reliable means of developing meaningful plans and monitoring those plans automatically.

### *1.5 Forms of Advanced Signal Control*

Intelligent Transportation Systems (ITS) tend to research areas of advanced signal control such as second and third generation control, fully adaptive traffic signal systems and even the smart highways. However, realistically in the United States, these systems are not ready for implementation and so less advanced systems are employed, such as time-of-day (TOD), due to factors like lack of funding to the transportation infrastructure *(5*). This research looks at improving and refining the current means of traffic signal plan procedures (TOD), which tend to be overlooked as areas of research. Existing literature focuses on the improvements that can be made by implementing advanced control methods with the resources provided by ITS. Until the more advanced methods become feasible in this country, it appears that little interest is taken in utilizing the ITS resources for less advanced signal development practices. This notion is reflected by the fact that no existing literature was discovered on the use of ITS data for enhancing TOD methods of signal development and implementation. The *Transportation Research Circular (6)* discusses research initiatives for advanced technology in traffic signal control systems because of the need for improvement in this area. These needs are due to the steady increase in traffic congestion, in some areas reaching crisis proportions, and the decreasing availability of land for the use of highway and road expansion *(6*).

There are four categories of traffic signal control:

- **First Generation**
- **1.5 Generation**
- **Second Generation**
- **Third Generation**

TOD signal control falls into the first generation category, which consists of selecting a timing plan from a library of stored plans, which have been developed off-line using a tool such as Synchro (5). 1.5 Generation is identical to first generation except that it has the automated ability to add plans to the library. Second and third generation control are advanced forms of control that implement traffic signal plans in real time based on existing traffic conditions. Third generation differs from second generation in that the cycle lengths and splits have the capability of variability, whereas second generation has fixed cycle lengths and splits *(5).* The U.S. is one of the few advanced countries in the world where adaptive control is not installed. This is due to the increased cost of surveillance for monitoring and maintaining the large number of detectors necessary for supporting the use of this type of control. Adaptive control reduces the need for timing plan updates and it handles incidents, holiday and special events more efficiently *(5*). These advanced forms of control are capable of using information about downstream traffic to update plan parameters at the upstream signals.

Minneapolis is one of the few cities in this country that is testing a second-generation adaptive control signal system. The project is described in the paper, *Addition of Adaptive Control to the Minneapolis Signal System: Issues and Solutions (8*). This project's aim is to serve as a representative model for medium-sized North American centrally controlled systems, which assesses costs, problems and potential gains from the addition of such a system. This project recognizes the fact that extensive detection inputs, beyond those installed for existing signal methods are needed to support advanced

forms of control. It is also addressing the many other issues to consider with advanced control. These include determining the operational status of the system, how to verify system requirements of the new system are being met, what sort of considerations must be met when adding a new system to an operating system, and many other issues involved with such an advancement *(8).* There are many challenges to be met before advanced forms of control are fully understood and supported affordably in this country.

### *1.6 Data and Data Collection*

Data is collected at signalized intersections in Northern Virginia by single inductive loop detectors. These metallic detectors are embedded in the roadway and produce a magnetic field. The metal of a car passing over the detector interferes with the magnetic field, thus permitting the detection of the vehicle by the detector. The single inductive loop detectors, referred to as system detectors, are recommended in the Traffic Control Systems Handbook to be placed 61 – 76 meters upstream from an intersection's stop-bar at a minimum (1). The northern Virginia system detectors are typically placed at approximately 100 meters upstream from intersection stop-bars. The placement of system detectors is a key consideration because lane discipline deteriorates in the vicinity of the intersection, especially during periods of spill back, and lane-changing maneuvers from upstream can produce significant errors in volume and occupancy readings. A system detector should never be placed where standing queues from the downstream intersection typically extend. Yet the detector should be placed close enough to the intersection to distinguish between vehicles that are using turning lanes rather than the through movement lanes. Single loop refers to the fact that only one detector is placed upstream from the intersection removing the capability of directly measuring speeds.

Volumes and occupancies are directly measured and the speed is an internal calculation formulated based on estimates of vehicle lengths and detector lengths. Thus, speed was not used in this research.

Volume is defined as the number of vehicles that pass over the system detector in a given time period. It is simply a count of the cars that is generally expressed in vehicles per hour (VPH) or in the case of the Northern Virginia research in vehicles per 15-minutes (VP15m). According to the Highway Capacity Manual, a typical roadway capacity for one hour is 1900 vehicles. Occupancy is defined as the percent of time a vehicle occupies a detector. Occupancies are reported as a percentage. Once occupancies reach 25%, the roadway can typically be classified as saturated. Saturation occurs when the volume to capacity ratio (V/C) is near to or greater than one. Occupancies greater than 25% lose meaning as they can fluctuate between 25% and 100% for varying values of volume, with no particular correlation other than the volumes are typically at least greater than 600 VPH at this point.

### *1.7 Data Screening Tests*

The data extracted from the database is cleansed prior to its use. Much "bad" data is returned from the detectors. There are many possible reasons for this, such as damaged or dead detectors. Cleansing the data allows the user to look only at reasonable data, thus removing many outliers and observations that will skew the clustering results. Screening rules were determined based on typical data relationships in the database. For this research, the screening rules were all used. The screening rules are as follows to remove bad data:

**Non-zero test:**
- Volume AND Occupancy AND Speed ≠ 0

**Prescreening test:**
- Volume AND Occupancy AND Speed >= 0
- Volume < 3100 AND Occupancy < 100
- Volume >= Occupancy

**Feasible Volumes:**
- IF Occupancy = 0 OR 1 THEN Volume < 580
- IF 1 < Occupancy <= 15 THEN 1 < Volume < 1400
- IF 15 < Occupancy < 25 THEN 180 < Volume < 2000
- IF Occupancy >= 25 THEN Volume > 500

The methods the screening tests use to scrub data can be grouped into two categories, threshold value tests and traffic flow theory tests. The 'Non-zero Test' uses a threshold value test, the 'Prescreening Test' uses both threshold value and traffic flow theory tests, and the 'Feasible Volumes' test uses only traffic flow theory tests (18). Threshold value tests limit data to within physically reasonable values based on characteristics of volume and occupancy. Traffic flow theory tests restrict data to feasible combinations of volume for a given occupancy.

All rules were established by examining data from 5 arbitrary intersections in Northern Virginia for periods of up to one month. The screening tests were then applied to various intersections to test the procedures. Values of volume for the hourly intervals are given in vehicles per hour (VPH), the unit of measurement used in the database. Occupancies are given as a percentage of time vehicles are located over a detector. Speed is not used in the traffic theory rules because it is derived from volume and occupancy, an assumed vehicle length and an assumed detector length, producing inaccurate data. Figure 3, Figure 4, Figure 5 and Figure 6 show how the "Feasible Volumes" tests were

derived from the data. Based on typical volume and occupancy relationships derived from the 5 intersections investigated as shown in the graphs below, the numerical values for the threshold value, data screening tests were developed.

**Figure 3. Verification of Feasible Volumes test 1**

**Figure 4. Verification for Feasible Volumes test 2**

**Figure 5. Verification for Feasible Volumes test 3**

**Figure 6. Verification for Feasible Volumes test 4**

### *1.8 Project Scope*

This project research will introduce the use of data mining tools for timing plan development based on system detector data. The proposed procedure will be conducted on a subset of a single corridor in the Northern Virginia arterial network. The corridor studied will be a piece of the Reston Corridor, consisting of 3 coordinated intersections and 15 system detectors. Figure 7, below shows the majority of the Reston corridor layout taken from a Synchro file, from which the subset corridor is taken for the case study. The timing plan development scheme will be based only on Monday through Friday for the entire 24-hour period. The system detector data will be acquired from system detectors, or single inductive loops, located in select lanes throughout the corridor. Volumes and occupancies collected from these system detectors and archived in an Oracle Database in the Smart Travel Laboratory at the University of Virginia will be used to conduct this research. The Virginia Department of Transportation (VDOT) supplies the data to the Smart Travel Laboratory (STL) at the University of Virginia, which is aggregated to 15-minute observations.

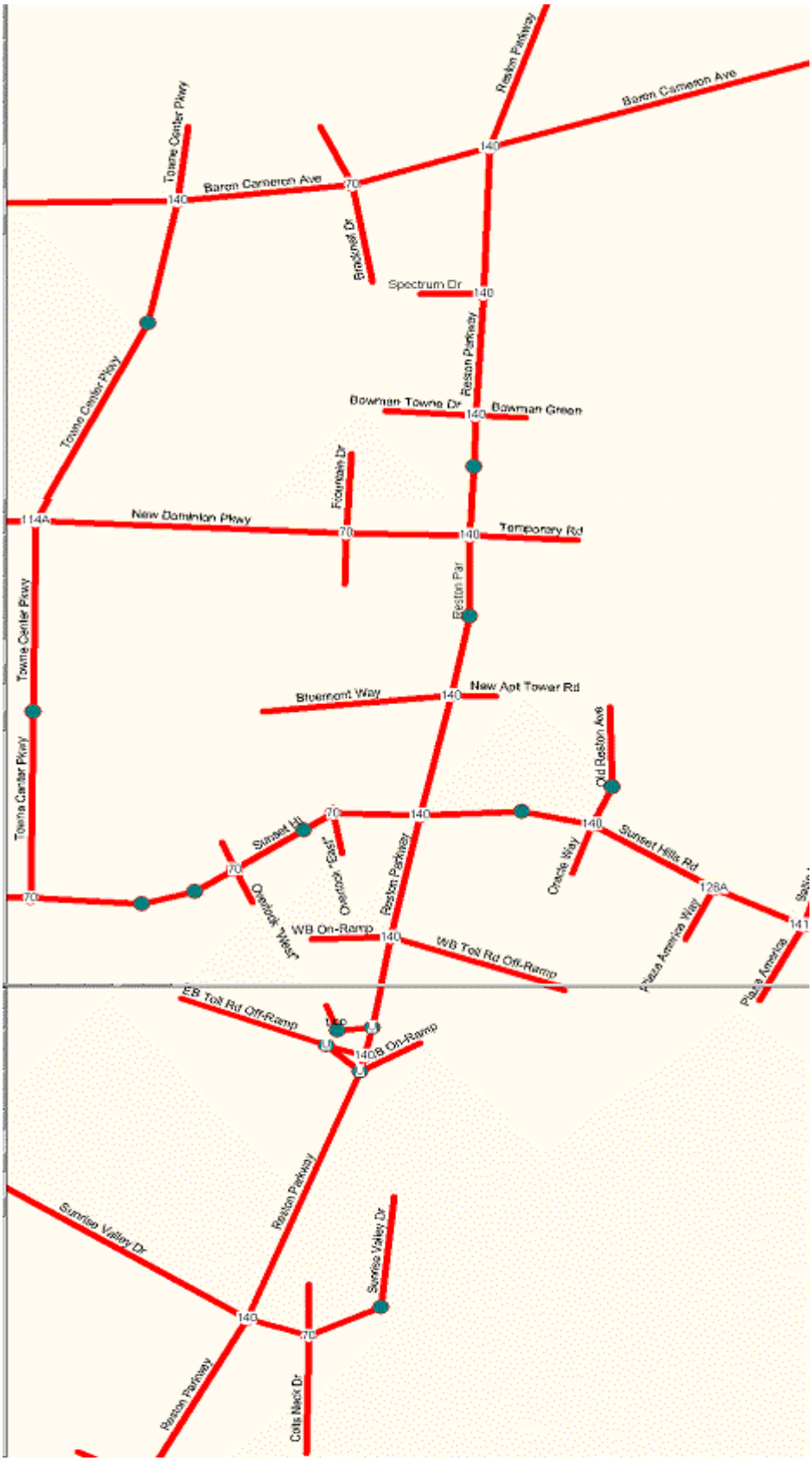



**Figure 7. Reston Corridor Layout**

### *1.9 Project Statement*

This project will contribute to the intelligent transportation systems (ITS) field by utilizing real-time detector data through the use of data mining tools to aid in the development of signal timing plans and fixed time-of-day (TOD) intervals for traffic signal plan implementation. The thesis is that data mining techniques, not traditionally used for timing plan development in transportation, such as clustering, can be used to improve the development of signal timing plans and fixed time-of-day (TOD) intervals for traffic signal plan implementations. The main objective of this project is to propose a procedure for utilizing detector data for improved plan development by detailing the following tasks:

- Use of data mining tools (cluster analysis) to extract information from a large database,
- Improve timing plans through use of data extracted from database versus the current method of one-day volume counts,
- Improve TOD intervals using cluster analysis on detector data with refined and expanded state definition, and
- Test clusters and plan performance through simulation and internal cluster validation.

Chapter 2 will provide background information on the signal timing plans and methods of current traffic control, while detailing current methods of timing plan development. Chapter 2 will also discuss related areas of research to the topics explored in this project. Chapter 3 will detail the problem formulation for each phase of the research, including the selection of a clustering method, validation of the clusters developed, timing plan development in Synchro and simulation with SimTraffic for plan evaluation. The proposed procedure will be outlined in detail in Chapter 4, fully

describing the tools used for this research and providing guidance in following and enhancing the procedure. Chapter 4 will provide the major deliverable of this project, the proposed procedure with guidelines for following the procedure and automating the procedure. Discussion of the results of the analysis based on a single corridor case study and a brief analysis at a single intersection in Northern Virginia will be introduced in Chapter 5. Evaluation of the proposed procedure and the applicability of this research are discussed in Chapter 6, with emphasis on the future research needs for a more robust procedure.

# Chapter 2. BACKGROUND

## *2.1 Signal Timing Plans*

The operation of a coordinated signal control system on an arterial corridor, or a series of signalized intersections operating under a common traffic signal plan, requires a timing plan for each signal in the corridor. A corridor-timing plan consists of four main elements: cycle length, splits, offsets and phase sequences *(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 signal *(1).*

Aside from the three main elements of a coordinated traffic signal as discussed above, there are many other components that must be taken into account in the development of timing signals. These components are as follows:

- Traffic Volume per lane movement
- Turn type (Protected or Permitted)
- Minimum Initial
- Minimum Split
- Maximum Split
- Total Split
- Yellow Time
- All-Red Time
- Lead/Lag
- Allow Lead/Lag Optimize?
- Vehicle Extension
- Minimum Gap
- Pedestrian Phase
- Walk Time
- Bus Blockages (#/hr)
- Heavy Vehicles (%)

- Growth Factor
- Peak Hour Factor
- Ideal Saturated Flow
- Lane Width
- Grade (%)
- Area Type
- Storage Length (ft)
- Storage Lanes (#)
- Right Turn on Red?

Clearly, the development of traffic signals is highly complex, which is why tools like Synchro are used (16). For this research, the Synchro files developed by VDOT were used so that the all the timing plan components listed above were already archived. The only alterations made to the Synchro files for this research are the volumes for the modified TOD intervals for which the timing plans are servicing. With the alteration of the volumes, Synchro optimizes the cycle length, split and offset to best suit the timing plan inputs.

### *2.2 Phase Movements*

Opposing movements at an intersection are defined by phases. Phases are numerical values (1,2,3,…8) assigned to through/right-hand-turn movements and left-hand-turn movements. Even phase numbers are always assigned to through/right-hand-turn movements and odd phase numbers are assigned to left-hand-turning movements. Figure 8 shows a sample intersection with phase assignments.

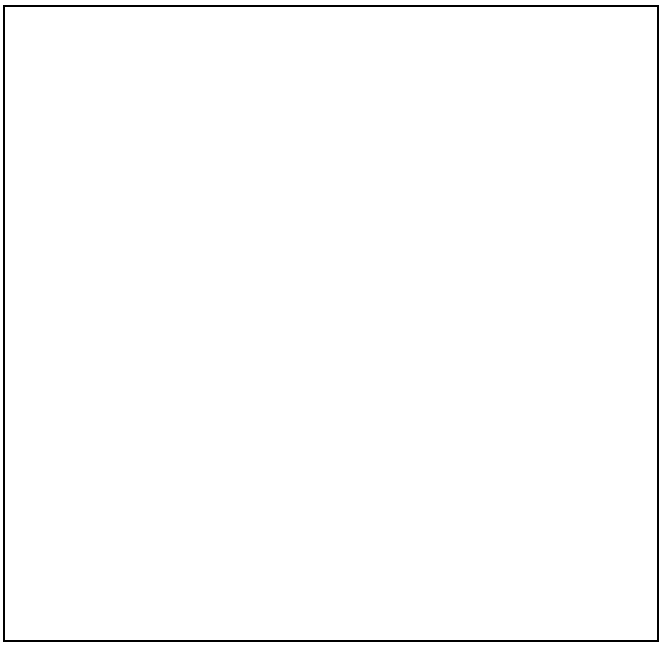


**Figure 8. Phase Diagram**

Phase numbers are always grouped together as shown in Figure 8. The only dynamic element of the phase diagram is whether phases 2/5 and 1/6 lie on the east-west direction or the north-south direction. This element is dependent on the direction of the main throughway. Phases 2 and 6 always correspond with the main throughway. If the main throughway lies in the east-west direction, then the phase diagram is as shown in Figure 8. If the main throughway lies in the north-south direction, then the phase diagram in Figure 8 would have to be rotated 90 degrees to the right.

### *2.3 Local Detection Control*

Traffic signals can be actuated, semi-actuated or pre-timed. Actuated signals are driven by the traffic conditions sensed by the local detectors. Local detectors do not collect volume, occupancy and speed data as do the system detectors. They are solely for the purpose of traffic signal actuation. Fully actuated control is extremely difficult to implement because of the difficulty and expense involved with maintaining enough

detectors to support the control. Second and third generation control run with fully actuated corridors. Vehicles trigger the detector to change the green split to that phase movement. Semi-actuation is the form of signal control used in the Northern Virginia arterial system. In semi-actuation, the main throughway is always given its preset green split time, even if no vehicles are detected at the intersection. However, the side streets will only maintain a minimum green time allotted to that phase split if vehicles are not detected. The remaining green time that would make up the full side street split is given back to the main throughway. If vehicles are detected on the side streets, then the maximum green time is given. Non-actuated control would exist at an intersection with no local detectors and the signal would operate under fully pre-timed signal parameters.

### *2.4 TOD Plan Methodology and Issues*

The most widely used method for timing plan selection and implementation 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 affective 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 a single volume count from the critical intersection. The volume count used for timing plan development in Synchro is taken from the traffic engineers' hand-counts of cars during assumed peak traffic time for the TOD interval. This single-day count is used for developing a timing plan for the entire corridor. Therefore, one will note that the challenge in designing a TOD system lies in identifying the appropriate time intervals for plans, and then developing effective corridor plans to operate within each interval. Another challenge

faced by traffic engineers is monitoring the performance of timing plans over time and retaining up-to-date timing plans. Because of the time and effort that goes into the current method of TOD plan development, the plans are generally left in place for many years, with no automated form of performance feedback. The use of electronic data and data mining tools would make automated timing plan development and maintenance readily available. Another issue that must be overlooked with the current means of TOD traffic control is that variance in traffic conditions can not be accounted for and variance over time may go overlooked until conditions become severe. Figure 9 and Figure 10 show a volume vs. time plot in the northbound and southbound direction for an intersection in the Reston corridor. Volume data from March 8, 2000 until September 29, 2000 were plotted. Traffic trends remain similar over such a short period of time, but there are erroneous days where variant traffic conditions get serviced by timing plans constructed for "normal" conditions. With automated maintenance tools, that will be made possible with the use of data mining tools, erroneous days and changing trends over time can be detected and archived. This will allow for the development of theories and rules based on traffic variance-time/event trends, thus preparing for changes in the future before they occur. The TOD itnervals may also change over time, where data mining tools would allow for detection of slight variations in transition times.

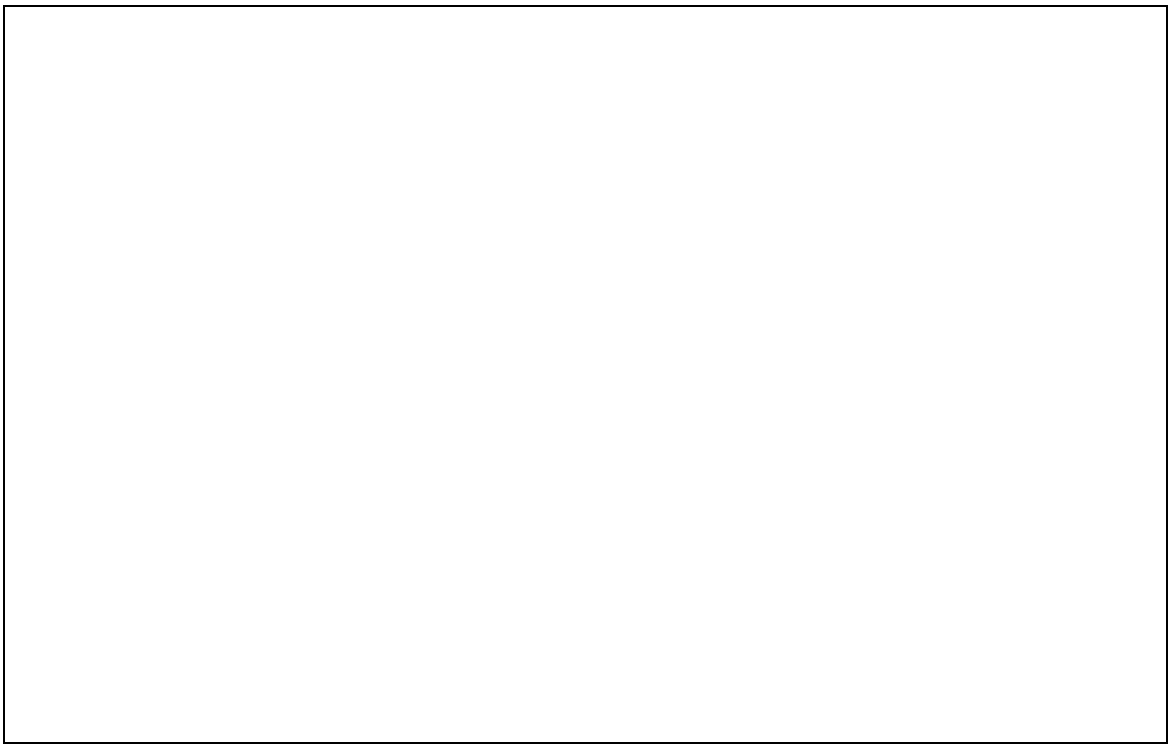


**Figure 9. NB Volume vs. TOD at one intersection**

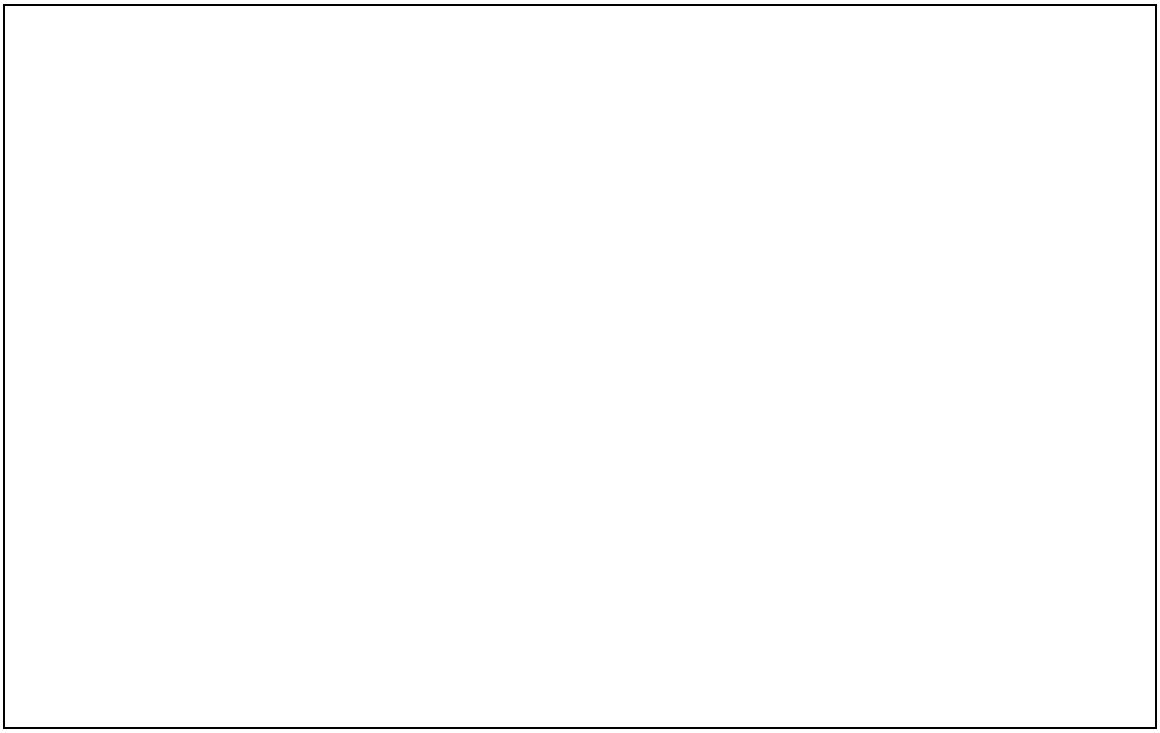


**Figure 10. SB Volume vs. TOD at one intersection**

### *2.5 Proposed State Definition*

Time-of-day (TOD) signal control is an example of a form of 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 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 a refined form of data available from the system detectors to identify TOD intervals.

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 each intersection in the corridor. The directional phase movements are identified by their corresponding phase numbers, which are denoted in Figure 8. In addition, to account for the difference in scale between volume and occupancy measures, the values were standardized using a Z-score, (Z), which represents a dispersion or spread from the mean that each value lies and is defined in the following equation.

$$\mathbf{Z = X - M / \sigma}$$

Chapter 5 will investigate alternate input cluster variables in the 'Sensitivity Studies' section in addition to equally weighted, standardized variables. Since volume and occupancy represent different traffic states, where occupancy values lie on a percent scale of 0 – 100 and volume values lie on a numeric scale of 0 – 1900+, the standardization process is necessary to transfer these values to a uniform, meaningful scale with no units (13). The possible effects of variable weighting will be discussed in more detail in Chapter 5, where consideration of un-standardized occupancy and volumes are taken into account. For the scope of this research, the detectors and cluster variables were 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. The state definition used is as follows, with each variable number assigned according to its phase number. Not all intersections have system detectors located at every phase, so the state definition may vary from intersection to intersection depending on the availability of system detectors.

**X(t) = (V1, O1, V2, O2, V3, O3, V4, O4, V5, O5, V6, O6, V7, O7, V8, O8),**

Where **X(t)** = system state at time t

**V1** = standardized phase 1 volume at time t (NBL)

**O1** = standardized phase 1 occupancy at time t (NBL)

**V2** = standardized phase 2 volume at time t (SB)

**O2** = standardized phase 2 occupancy at time t (SB)

**V3** = standardized phase 3 volume at time t (EBL)

**O3** = standardized phase 3 occupancy at time t (EBL)

**V4** = standardized phase 4 volume at time t (WB)

**O4** = standardized phase 4 occupancy at time t (WB)

**V5** = standardized phase 5, volume at time t (SBL)

**O5** = standardized phase 5, occupancy at time t (SBL)
**V6** = standardized phase 6 volume at time t (NB)
**O6** = standardized phase 6 occupancy at time t (NB)
**V7** = standardized phase 7 volume at time t (WBL)
**O7** = standardized phase 7 occupancy at time t (WBL)
**V8** = standardized phase 8 volume at time t (EB)
**O8** = standardized phase 8 occupancy at time t (EB)

## *2.6 RELATED RESEARCH*

Data mining tools are not widely used in transportation systems (7). In fact system detector data collection is a fairly recent advancement with the rise of ITS and has not yet been utilized to its full capacity. Traffic may be viewed as unpredictable and uncontrollable, but with archived data that is now available, it can be shown that traffic is in fact predictable to a degree and control can be improved with the utilization of this data. There are other DOT's that have looked into advanced forms of control such as traffic responsive and second generation, where the system detector data is necessary to support such control techniques, but it has not been found to be used for TOD signal control (6). Data mining tools are useful for uncovering patterns in data and making classifications and these notions can be highly beneficial in transportation systems. These data mining techniques have been used in many other fields and areas to produce similar results from many types of data sets.

### 2.6.1 Data Mining as an Emerging Field

Data mining is utilized in the disciplines of computer science and statistics and is making progress in extracting information from large databases (20). It is an emerging field that promotes the progress of data analysis. Due to the competitive nature of today's business

economy, information technology has been invested in heavily to aid in the management of effective business performance. Over the last three decades, increasingly large amounts of critical business data have been stored electronically and this volume is expected to continue to grow considerably in the future (20). Despite this wealth of data, many companies have been unable to fully capitalize on its value. This is because the information implicit in the data is not easily discernable without the use of data mining tools. Data mining tools allow businesses to leverage their data effectively and obtain insightful information that can give them a competitive edge. It enables them to discover previously undetected facts present in the data.

Data mining tools can provide benefits to any number of potential users. The finance and insurance industries have long recognized these benefits, but these principles can be applied in many areas. For example the retail/marketing sector, the banking sector, the insurance and health care sector, the transportation sector and the list goes on to those who can reap benefits from data mining tools (20). The following list summarizes some of the benefits that each of these sectors can achieve (20).

*Retail/Marketing*

- Identification of buying behavior patterns from customers
- Finding associations among customer demographic characteristics
- Prediction of customers responsive to mailing

*Banking*

- Detection of patterns of fraudulent credit card use
- Identification of “loyal” customers
- Prediction of customers that are likely to change credit card affiliation
- Determination of credit card spending by customer groups
- Finding hidden correlations between different financial indicators
- Identification of stock trading rules from historical data market

*Insurance/Health Care*

- Claims analysis – determination of which medical procedures are claimed together
- Prediction of which customers will buy new policies
- Identification of behavior patterns of risky customers
- Identification of fraudulent behavior

*Transportation*

- Determination of distribution schedules among outlets
- Analysis of loading patterns
- Identification of seasonal and time-of-day traffic trends
- Location of high risk incident areas

There is an extensive body of technology that exists and continues to evolve that can be used to construct data mining functions. A number of data mining methods exist that can be classified in four major groups: Associations, sequential patterns, classifiers and clustering (20). In associations, a collection of items and a set of records, each of which contain some number of items from the given collection exist for which an association function is established which returns affinities that exist among the collection of items. For example, these affinities can be expressed by rules such as "72% of all the records that contain items A, B and C also contain items E and F." With sequential patterns, a transaction log exists, which identifies transaction and product information, generally without customer identity. A sequential pattern function will analyze collections of sets of products a customer buys in every purchase order. With classifiers, there exists a set of records with a number of attributes, a set of tags (representing classes of records) and an assignment of a tag to each record. A classification function examines the set of tagged records and produces descriptions of the characteristics of records for each of the classes. These class descriptions can be used to tag new records. In clustering, there exists a set of untagged records. Since no classes are known, it is the

goal of a cluster to produce a reasonable segmentation of the set of input records according to some criteria. These data-mining operators can be used cooperatively or individually. With automated techniques as those described above, businesses can utilize the database information to discover trends and improve on current practices.

### 2.6.2 Cluster Analysis Applications

Cluster analysis deals with automating a commonly utilized human activity of forming classes or groups of similar objects. The objects to be clustered could be of any origin, from hospital patients, product brands and insect species to traffic data. Cluster analysis has been widely used in many diverse disciplines such as biology, psychology, archaeology, geology, marketing, information retrieval, and remote sensing (12). Clustering in computer science and engineering has been a more recent outcome solving many problems with pattern recognition and image processing. In these fields it has been used for things such as unsupervised learning, speech and speaker recognition, work-load characterization, crime detection and image registration. Cluster algorithms may be applied in many different fields to many different domains, but for all, the outcome is a grouping of underlying themes in a data set that may not be intuitive or easily established without such a tool.

Francois-Joseph Lapointe and Pierre Legendre performed a research project using hierarchical cluster analysis at the University of Montreal to distinguish between different types of single malt whiskies (17). The data consisted of Scottish produced single malt whiskies totaling to 300 varieties. Single malts differ in nose, color, body, palate and finish. To produce a connoisseur's guide to Scottish malt whiskies, they had to be distinguished base on three major questions:

1. What are the major types of single malt whiskies that can be recognized and what are their chief characteristics and best representatives?
2. What is the geographic component in that classification?
3. Do the various categories of characteristics – nose, color, body, palate and finish – lead to the same classification?

The first and third questions will be of interest here because the whiskies will be categorized based only on the variables; nose, color, body, palate and finish, using hierarchical clustering. These questions can be answered with the results of cluster analysis. The geographic components will be used for checking the clustered classification a posteriori and determining whether location effects the categorized whiskies. By comparing raw data sets (canonical analysis), distance matrices (correlation matrices) and dendrograms (consensus measures), these questions can be answered. This process is similar to that done in this project, where groupings were found in the data and the clusters formed were validated using methods similar to those mentioned above.

A distance matrix was constructed for the malts based on color, nose, body, palate and finish, where each description was scored in such a way that the relationships could be represented numerically. The clustering used Ward's minimum variance method, detailed in *Section 3.4*, to form a dendrogram for depicting the whisky clusters. A cophenetic matrix, described in *Section 3.10.1.5*, was computed from this dendrogram, in which the distances between objects is equal to the value of the fusion level where these two object were joined to the same cluster. The distance matrix from the a priori canonical analysis was compared to the dendrogram of the cluster analysis, where the null hypothesis tested is that the two comparisons are no more similar than randomly

generated dendrograms with the same number of objects, random topology and random labels would be. When the resulting classification of single malt whiskies was compared to geographic locations, it was shown that the whiskies could not only be characterized by physical properties, but also by distillery traditions and regions, where they are effected by soil, water, temperature, etc. This research not only classified single malt whiskies by defining characteristics and regions, but also characterized the whiskies based on clustered characteristics. The performance of comparisons among raw data, distance matrices and dendrograms was used to validate clusters.

The validation of the clusters formed in the single malt whiskey example and the idea of forming logical groups based on data characteristics where no response variable exists follows the idea of the research being done on volume and occupancy traffic data.

# Chapter 3. PROBLEM FORMULATION

## *3.1 Cluster Tools and Algorithms*

SAS, a software system for data analysis, was the main tool used to implement data mining procedures, clustering in particular, in this research (10). The cluster analysis was done with the SAS software using 15-minute volume and occupancy data obtained from an Oracle database in the Smart Travel Laboratory. This data is based on a continuous, quantitative scale. The purpose of cluster analysis is to place objects into groups or clusters suggested by the data, not defined a priori, such that objects in a given cluster tend to be similar to each other and objects in different clusters tend to be dissimilar (9). A vast number of clustering methods have been developed in several different fields, with different definitions of clusters and dissimilarity among objects. The choice of clustering algorithm depends both on the type of data available and on the particular purpose (13). Since cluster analysis is used as a descriptive or exploratory tool unlike statistical tests, which are used for confirmatory purposes, it is permissible to choose a clustering method based on cluster runs from the same data set. Thus suggesting and testing the theories introduced in this project with a clustering algorithm is sufficient in providing information on what the data are indicating.

The majority of clustering methods in the classification literature fall into one of two types of cluster algorithms; disjoint (partitioning) or hierarchical methods (13). Disjoint clusters place each object in only one cluster, where the number of clusters, k, have been defined a priori. Hierarchical clusters are organized so that one cluster may be entirely contained within another cluster, but no other overlap between clusters is allowed and the clusters are joined from n observations until only one cluster remains. SAS

procedures for clustering are oriented toward disjoint or hierarchical clusters from coordinate data, distance data, or a correlation or covariance matrix.

For the recommended procedure, hierarchical clustering has been used, however the optimal clustering method was not fully researched and may be further investigated for future timing plan development procedures. Hierarchical clustering seems advantageous to disjoint clustering (13). Disjoint clustering uses a k-means method where the number of clusters to be formed must be pre-determined. The determination of the optimal number of timing plans before the cluster analysis results in uncertainty and error, since the number of timing plans or clusters is an unknown statistic that cannot be firmly established prior to the cluster analysis. The disjoint clustering algorithm would have to be run several times with different values of k to retain the clustering that appears to provide the most meaningful interpretation based on data characteristics or graphs (13). Thus, hierarchical clustering appears to be the best choice for the purpose of supplying the data necessary for determining the optimal number of clusters based on the cluster analysis. Another disadvantage with disjoint clustering includes the non-stability of the clusters formed due to the choice of initial cluster seeds, which are affected by the order in which the data are read into the computer. Because of the large number of choices for the number of clusters and the location of the cluster seeds associated with each cluster, this procedure may become computationally infeasible, especially with large data sets (10). Hierarchical cluster analysis, on the other hand, results in much more stable clusters due to the procedure implemented where each observation begins in a unique cluster. Clusters are then joined based on the minimum dissimilarity measure between clusters, until only one cluster remains.

### *3.2 Introduction of Research Case Studies*

Two exploratory case studies were performed to test the procedure proposed in this project. This proposed procedure offers a method for TOD timing plan development that creates TOD intervals and timing plans for those intervals based on clustered system detector data. This procedure introduces methods of timing plan development with the capability to greatly reduce the time spent on plan development, while creating better suited TOD plans for current traffic conditions. This process also provides a means to automating the plan development and maintenance process.

The first case study is performed on a series of three intersections to form a small corridor for testing the validity and methods of the proposed procedure on a coordinated arterial. Figure 12 shows the lane configuration for this corridor, which is a subset of the full Reston corridor. This subset corridor consists of New Dominion, Bluemont and Sunset Hills, all intersecting with Reston Parkway. This small corridor will support theories proposed for the procedure with a fairly simple data set. The second case study consists of a single intersection at Baron Cameron and Reston Parkway. Figure 11 shows the lane configuration of this intersection. This will look at a simplified version of the procedure with a small data set taken from seven months of historical data from one intersection. The procedure will be evaluated based on single intersections supported by the Baron Cameron and Reston Parkway intersections to draw conclusions on the performance of the plan on single intersections.

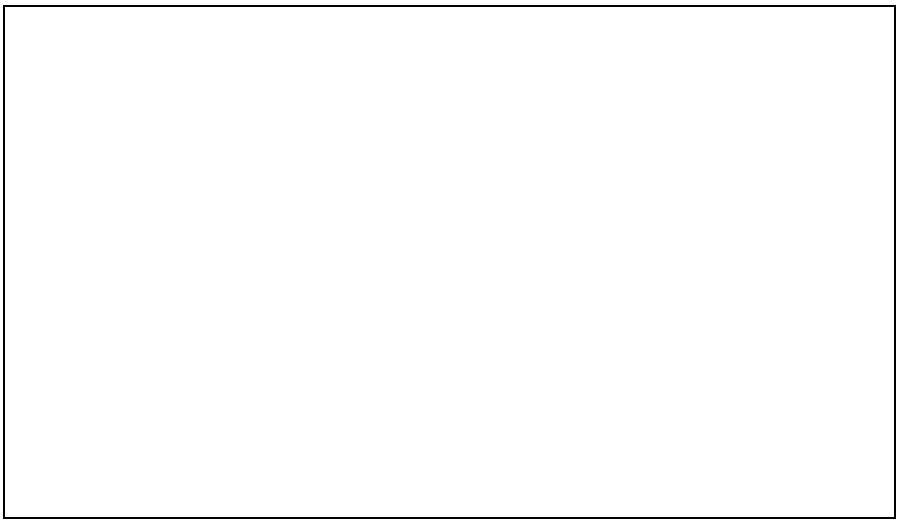


**Figure 11. Reston - Baron Cameron Intersection Layout**

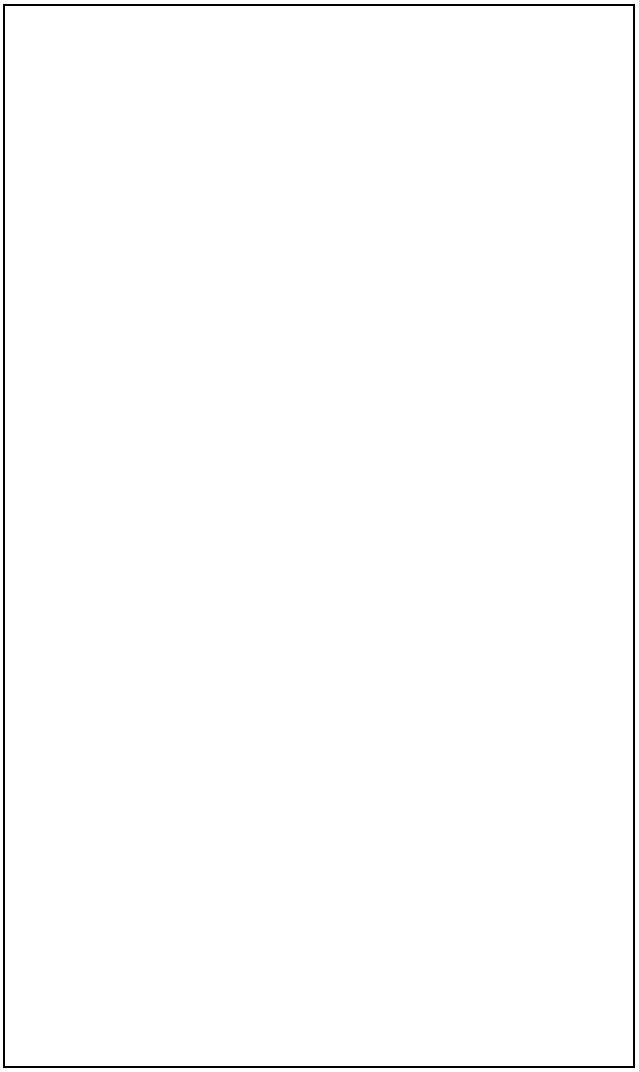


**Figure 12. Reston - Sunset Hills, Bluemont, New Dominion Intersections Layout**

### *3.3 Hierarchical Clustering*

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, 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 when categories of the data are not defined a priori. Hierarchical 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 for which timing plans should operate in based on similar traffic characteristics.

With Hierarchical clustering, each observation begins in a cluster by itself. The two closest clusters are merged to form a new cluster that replaces the two old clusters. Merging of the two closest clusters is repeated until only one cluster is left. At each level of the merging process, there exists one less cluster due to the joining of a cluster from the previous level (13). The various clustering methods differ in how the distance between two clusters is computed.

### *3.4 Cluster Methodologies*

There are many clustering methods that can be implemented for a cluster analysis. These include methods such as average linkage, Ward's minimum variance method, centroid, complete linkage, single linkage, and density linkage (10). In average linkage, the

distance between two clusters is the average distance between pairs of observations, one in each cluster. Average linkage tends to join clusters with small variances and is slightly biased toward producing clusters with the same variance (10). In Ward's minimum variance method, the distance between two clusters is the ANOVA sum of squares between the two clusters added up over all the variables. At each generation, the within-cluster sum of squares is minimized over all partitions obtainable by merging two clusters from the previous generation. Ward's method tends to join clusters with a small number of observations and is strongly biased toward producing clusters with roughly the same number of observations (10). It is also very sensitive to outliers. In the centroid method, the distance between two clusters is defined as the squared Euclidean distance between their centroids or means. The centroid method is more robust to outliers than most other hierarchical methods (10). In complete linkage, the distance between two clusters is the maximum distance between an observation in one cluster and an observation in the other cluster. Complete linkage is strongly biased toward producing clusters with roughly equal diameters and can be severely distorted by moderate outliers. Complete linkage (furthest neighbor) determines the distances between clusters by the greatest distance between any two objects in the different clusters (10). This method is inappropriate if the clusters tend to be elongated or of a chain nature. In single linkage, the distance between two clusters is the minimum distance between an observation in one cluster and an observation in the other cluster. Single linkage sacrifices performance in the recovery of compact clusters in return for the ability to detect elongated and irregular clusters. It tends to string objects together in cluster formation (10). Density linkage encompasses the $k^{th}$-nearest neighbor method, the uniform kernel method and the Wong's hybrid

method. Wong's hybrid clustering method uses density estimates based on preliminary k-means or disjoint clustering. The $k^{th}$-nearest neighbor method uses $k^{th}$-nearest neighbor density estimates and the uniform-kernel method uses uniform-kernel density estimates. These density linkage methods do not apply constraints to the shapes of the clusters and, unlike most other Hierarchical clustering methods, are capable of recovering clusters with elongated or irregular shapes (10). Yet density linkage is less effective at recovering compact clusters from small samples.

Studies have been done comparing the various methods of cluster analysis. Many of the methods are biased towards finding clusters possessing certain characteristics related to size, shape or dispersion (9), (10). For instance, Ward's minimum variance method and k-means tend to find clusters with roughly the same number of observations in each cluster. Average linkage tends to be biased towards finding clusters of equal variance. Many clustering methods tend to detect compact, roughly hyper-spherical clusters and are incapable of detecting clusters highly elongated or irregular shapes. The methods with the least bias are those based on non-parametric density estimation such as single linkage and density linkage.

### *3.5 Suggested Cluster Methodology*

The appropriate clustering method was only briefly investigated; however, based on studies done by Milligan and cluster comparisons from the brief review, an appropriate method was selected for use based on data characteristics and preliminary results (9). The outputs of the centroid, Ward, K-nearest neighbors density and single linkage methods were tested and compared, based on the data characteristics these methods utilize and the capabilities of the methodologies. The other methods were ruled out based

on performance observations by Milligan, such as the severe distortion of the complete linkage method by moderate outliers and its inability to detect elongated clusters. Each method was tested on the same data set, with five clusters being formed with each of the methods. Cluster outputs for these analyses are shown in Figure 13, Figure 14, Figure 15 and Figure 16.

The density method was ruled out due to the inability of the method to detect clusters with large enough number of members comprising the clusters. Figure 13 portrays the problem with clustering the volume, occupancy data with a density method. Different numbers of K were chosen for the nearest neighbor value and the results did not change. According to studies, density linkage methods do not apply constraints to the shapes of the clusters and do not perform well at recovering compact clusters from small samples. Since the volume, occupancy data sets used for the purpose of timing plan development are fairly small samples, density methodologies seem to be inappropriate.

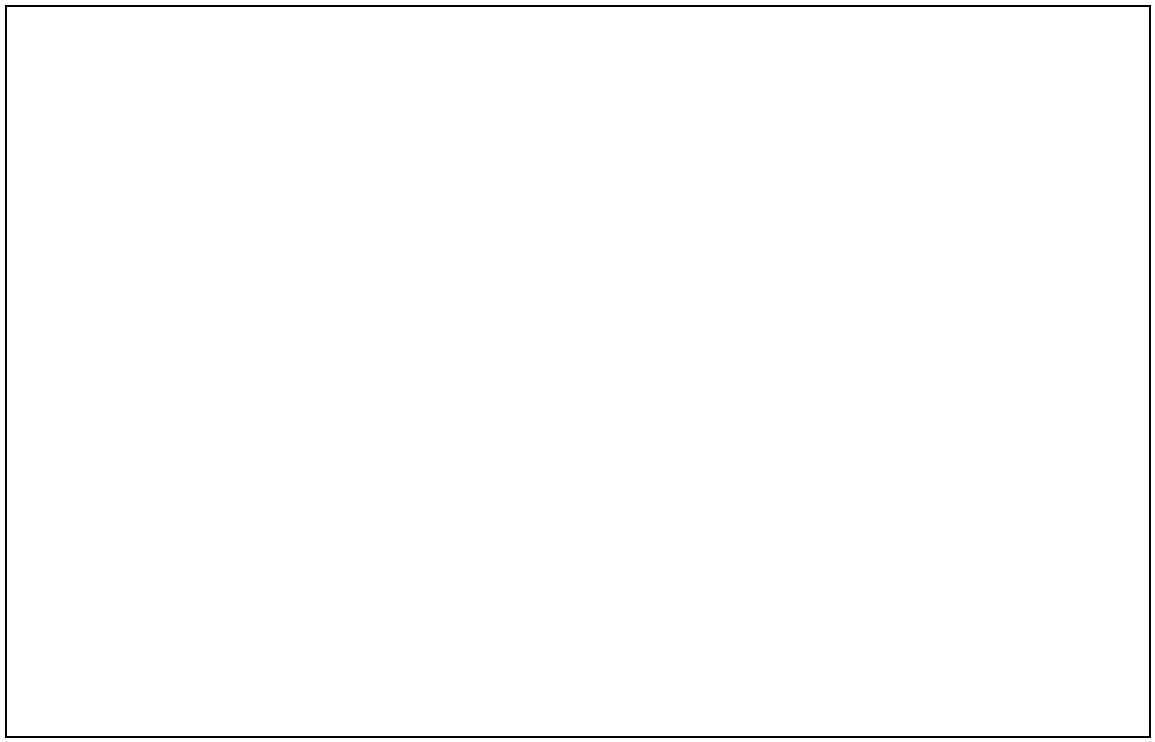


**Figure 13. Cluster vs. TOD Results for K-nearest Neighbors Method**

The single linkage method has been ruled out due to its inability to place a 'minimum number of observations' constraint on the clusters. This results in clusters being formed based only on a few number of observations, which could cause an inefficient transition between timing plans due to the shortened duration in each timing plan. See Figure 14. It also does not uncover compact clusters from data sets, rather it finds highly elongated clusters. Since the volume and occupancy data are quite similar, a method that uncovers compact clusters is preferable.

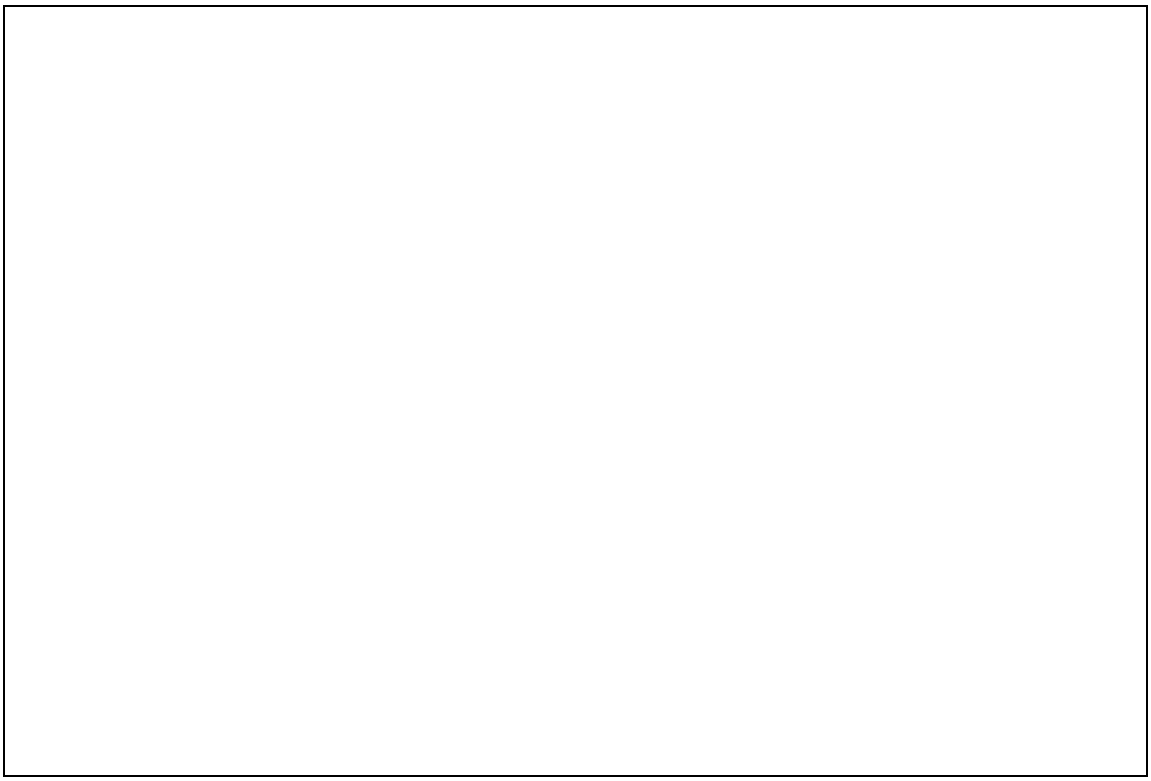


**Figure 14. Cluster vs. TOD Results for Single Linkage Method**

The centroid method produces fairly good clusters. Figure 15 illustrates this with the intuitive TOD intervals that form the clusters. Cluster 1 can be classified as the post-AM / post-PM period. It is based on observations from times that range from 9:15 – 11:30 and 19:00 – 19:45. It makes sense that traffic conditions occurring at these two times, directly after the two largest peak periods in a day, should have similar volume, occupancy pairs. Cluster 2 represents an off-peak period ranging from 22:15 – 5:30. Cluster 3 can be classified as a pre-AM / pre-Off-peak period ranging from 5:45 – 6:45 and from 20:00 – 22:00. Again, these TOD intervals that make up cluster 3 make sense in that similar traffic conditions occur at these alternate times. Cluster 4 covers the AM peak period and most of the lunch and mid-day period. The times range from 7:30 – 9:00 and approximately 11:45 – 17:00. This particular data set appears to be missing

observations from the PM peak period, a situation addressed in the ‘Sensitivity Analyses’ section of chapter 5.

**Figure 15. Cluster vs. TOD results for Centroid Method**

Ward’s method produces very similar results as the centroid method. See Figure 16. Centroid and Ward’s methods seem more appropriate than the others for this data because they are able to produce clusters based on a constraint for a minimum number of observations to exist in each cluster formed. This is an important constraint due to the necessity of cluster formations with a large enough number of observations to support a timing plan for an appreciable period of time, to be developed for that cluster. For the case of the volume and occupancy data, it is not necessary to choose a method that will detect irregular or elongated clusters. It is preferable that the clusters maintain a nearly hyper-spherical shape to ensure that cases that should operate in opposing timing plans do not get placed in the inappropriate cluster, due to the similarities in variables of certain

opposing timing plans. For example, northbound and westbound volumes and occupancies may be very similar for the AM and PM peak periods, while the main differences lie in the southbound and eastbound variables.

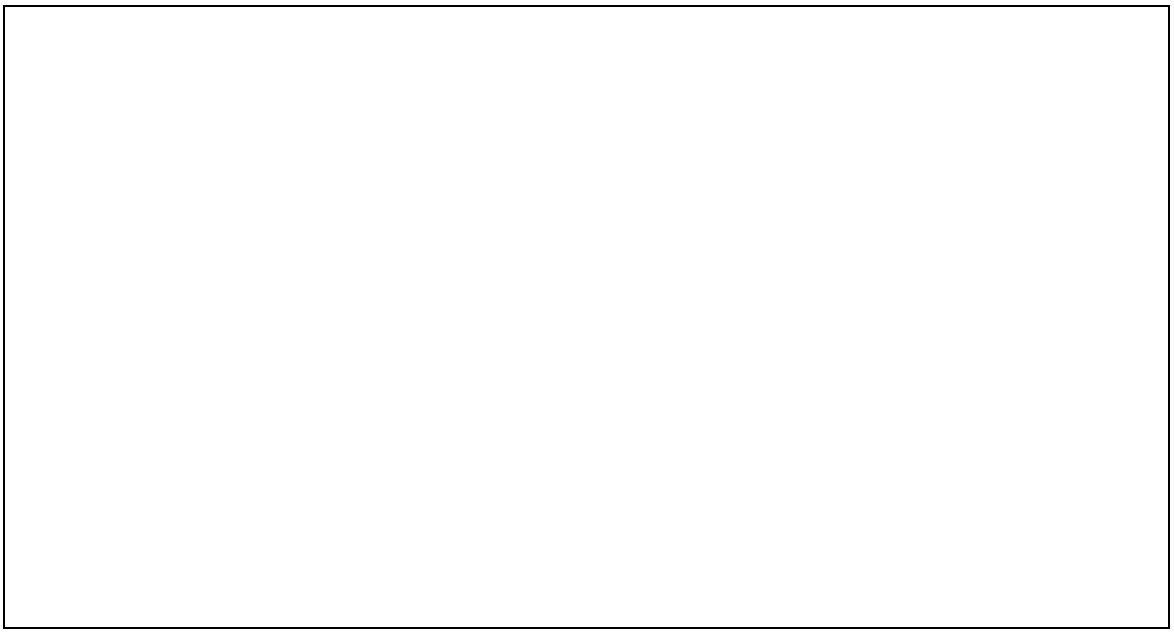


**Figure 16. Cluster vs. TOD Results for Ward's Method**

The only outstanding difference between Ward's method and the centroid method is that centroid is robust to outliers, whereas Ward's is not. Thus, for the purpose of this research, the centroid method will be used with the recommendation that further analysis may be conducted in the future for insight into the most beneficial methodology for clustering traffic data into timing plan intervals. Since outliers are inherent to traffic data due to things like incidents and holidays, the cluster method should not be overly sensitive to outliers. Figure 17 shows where the volume and occupancy centroids of each cluster lie for the centroid cluster methodology. The error bars represent the standard deviation within each cluster. The centroids in Figure 17 are intuitive for their corresponding timing plan periods. For instance, cluster 2 represents the off peak TOD

interval and the smallest volume and occupancy represent this cluster centroid. It is clear that the fifth cluster (C5) has a much larger occupancy mean than the other clusters. This is an issue that has arisen in preliminary analysis with the formation of bad clusters, discussed in the following section. Overall, this figure demonstrates that the clusters formed by the centroid method are meaningful for TOD interval plans. Refer to the internal cluster validation in *Section 3.10.1* for further testing of the formation of clusters under the Centroid methodology.

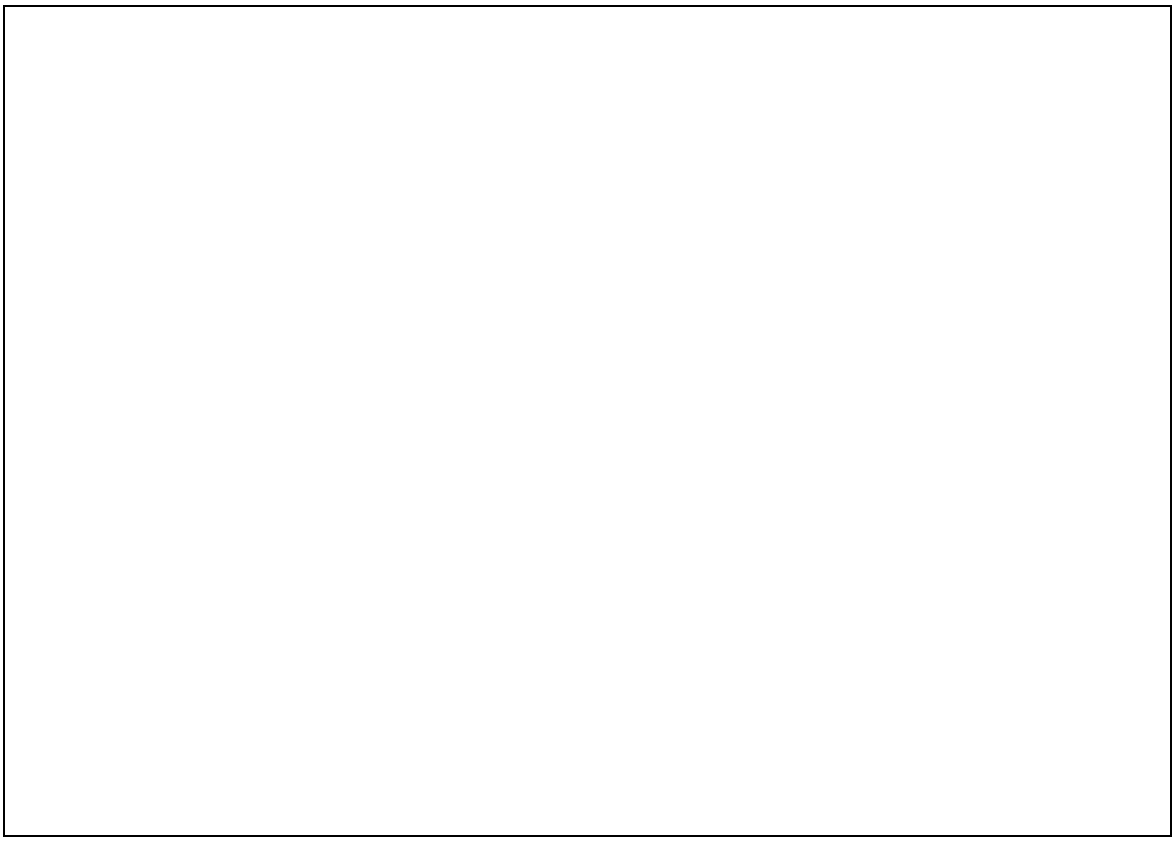


**Figure 17. Centroid Cluster Centroids and Standard Deviations**

### *3.6 Interpreting "Bad" Clusters*

The centroid and Ward's methods produce fairly good clusters as can be seen in Figure 15 and Figure 16. The times of day for which the clusters are formed from make sense intuitively. Situations arise in cluster analysis where "un-clean" clusters may be formed that don't follow an intuitive TOD scheme as do the majority of the clusters formed. This is apparent in Figure 15 and Figure 16 for cluster 5, which doesn't fall into an intuitive time interval for a timing plan. These random clusters arise in all of the methodologies and are usually due to one or two extreme variables existing for some of the observations. Data cleansing tests are induced to alleviate most such cases from arising, but bad clusters can still be formed. For example cluster 5, mentioned above, was formed based on extremely large phase 6 occupancies, in the above 50% range. Since occupancies greater than 25% mean the roadway is saturated, it is rare to see occupancies greater than 25%. During peak periods, occupancies typically exist around the 20% - 25% region. The fact that these occupancies exist only on one phase during these times may clue the traffic engineer to the fact the roadway in the phase 6 direction just can't handle the volume of traffic during those times and that an additional roadway may be needed. In cases such as the above example, it would be preferable to discard cluster 5 as a timing plan, especially since there is not a clean TOD interval for that cluster. Instead, such results could be used for alerting and making recommendations to the traffic engineers for problems on the roadway and the need for possible physical alterations. Methods of dealing with bad clusters will be addressed in Chapter 5, for sensitivity analyses with cluster input variables.

### *3.7 Euclidean Dissimilarity Measure*

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 this case the squared Euclidean distance, which is the default measure in SAS for the centroid methodology. The squared Euclidean distance is as follows:

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

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 so that timing plans are not scattered individually throughout the day. The time lost due to transition between timing plans is not thoroughly investigated in this research and should be examined in future research. However, some insights into effects on performance due to transition are evident from simulation, which simulates over plan transition periods.

To illustrate that Euclidean distance measure is a viable measure of dissimilarity for the clustering procedure, a demonstration on a small data set was done. Twenty-nine observations were chosen randomly to cover a 24-hour period. The data set had been clustered prior to the demonstration, thus assigning each observation a cluster membership to be used in illustrating the validity of the clusters based on the Euclidean distance. The 29 observations came from four clusters representing an AM-peak, a Mid-day peak, a PM-peak and an Off-peak. A dissimilarity matrix was constructed from the standardized volume, occupancy pairs making up this small data set. The Euclidean distance was the measure of dissimilarity between observations used to construct the matrix. Figure 1 illustrates the Euclidean dissimilarity measure on the y-axis between each of the four clusters, represented as four unique series on the graph, and each observation in the data set. The 29 observations are labeled on the x-axis by their cluster membership. The graph shows that the dissimilarity tends to be the lowest between observations belonging to similar clusters. Observations in clusters 1 and 3 are less than 1 unit apart from each other and clusters 4 and 6 are comprised of observations between 1 and 2 units apart. Observations between opposing clusters are between 2 and 5 units of distance apart from each other. This is apparent in Figure 18 demonstrating that Euclidean distance does cluster observations in an appropriate manner for minimizing the within-cluster distances and maximizing the between-cluster distances.

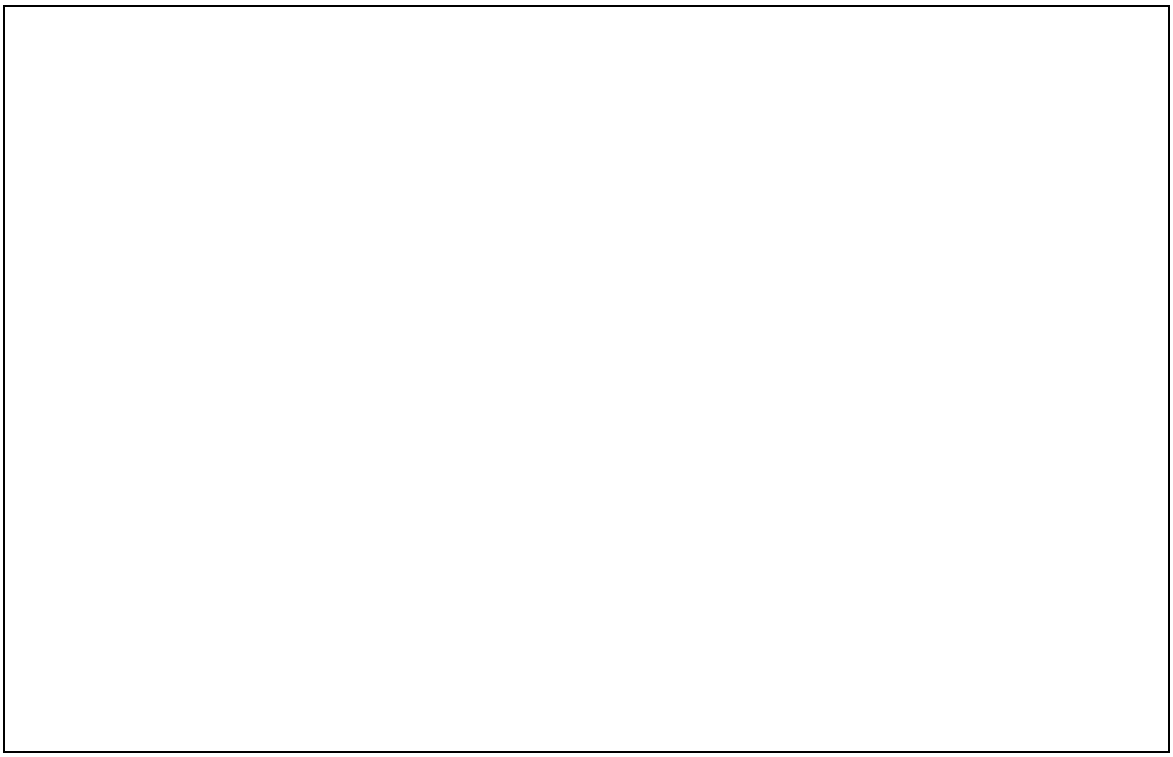


**Figure 18. Observation dissimilarity demonstration**

### *3.8 Determination of the Optimal Number of Clusters*

It is of importance with hierarchical clustering to determine the appropriate 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 that determine the optimal number of clusters are called "stopping rules." Statistics for determining the number of appropriate clusters are also numerous. Common statistics used for making such determinations include $R^2$ values, analysis-of-variance F-tests, the determinant of the within-cluster sum of squares matrix $|\mathbf{W}|$, the cubic clustering criterion (CCC), pseudo F statistic and pseudo $t^2$ statistic (9). For the traffic signal control procedure, the CCC will be recommended because it can be used with any of the hierarchical or disjoint clustering methods. To ensure a robust measure, the pseudo F and $t^2$ statistics will also be used in correspondence

with the CCC. The level of perfect replication from random partitions of an original data set will be investigated as a cluster analysis stopping rule tool from more recent research studies (14).

The $R^2$ value is only of use when the purpose is not to uncover real clusters or when the clustering method is average, centroid or Ward. Ordinary significance tests such as the ANOVA F-test are not valid for testing differences between clusters since clustering methods attempt to maximize the separation between clusters, thus drastically violating assumptions for normal significance testing. It cannot be assumed that the clusters are formed based on random assignment of observations to clusters, since that would defeat the purpose and methodologies of cluster assignment. The $|\mathbf{W}|$ criterion is an extremely conservative test because the cluster procedures in SAS attempt to minimize the trace of **W** rather than the determinant. There are alternate means of determining the number of clusters, but these are generally restricted to use with individual cluster methods. For instance the $k^{th}$-nearest neighbor clustering method can provide information for the number of clusters based on estimated number of modes versus k-values.

Since the cubic clustering criterion (CCC) can be used universally with all clustering methods and is a fairly accurate measure of determining the number of clusters, it will be recommended for use with the pseudo F and $t^2$ statistics for determining the number of clusters. According to stopping rule studies done by Milligan and Cooper, the CCC performed at a competitive rate, as the $6^{th}$ best, with the other 29 stopping rules tested (9). The stopping rules were tested based on prior knowledge of the correct number of clusters. The CCC does exhibit a fairly high rate of determining too

many clusters, but it did produce a very low number of solutions with too few clusters. The overall rate of correct determination of the true number of clusters for the CCC in Milligan and Cooper's studies was 74.3%. This was the sixth best overall rate of the 30 stopping rules investigated. The CCC is based on the assumption that a uniform distribution on a hyper-rectangle will be divided into clusters shaped roughly like hypercubes (9). In large samples, this assumption proves to give very accurate results. In other cases, the approximation is generally conservative.

The pseudo F and $t^2$ statistics will also be recommended as a measure of the appropriate number of clusters. Since SAS outputs these values for the cluster methodologies suggested above as well as the CCC, it will be advised that all three statistics are used together to choose the number of clusters. This should increase the rate of appropriate cluster choice that would not be achieved with using just one or the other. In an adaptation of the SAS User's guide (1990) and Sarle and Kuo (1993), it is recommended to look for consensus among these three statistics (15). In other words, local peaks of the CCC and pseudo F statistic combined with a small $t^2$ where a larger $t^2$ value occurs at the next cluster fusion. These criteria are most appropriate for compact or slightly elongated clusters, preferably clusters that are roughly multivariate normal (15). The pseudo F and $t^2$ statistic are also related to stopping rules tested by Milligan and Cooper. Calinski and Harabasz developed the pseudo F statistic. The pseudo $t^2$ statistic can be transformed from Duda and Hart's test statistic: J_e (2) / J_e (1) (15). According to the stopping rule study by Milligan and Cooper, the Calinski and Harabasz rule performed the best overall of the 30 rules tested with a 90.3% rate of the correct level of clusters (9). The Duda and Hart statistic performed the second best in the 30-rule test

with a rate of 89.8% correct level of clusters. Both of these stopping rules tend to suggest one too few clusters, which is where the majority of the mis-classification lies.

Combining the pseudo F and $t^2$ statistic with the CCC, whose main error lies in producing too many clusters, should balance each other, providing fairly reliable determinations for cluster number.

### 3.8.1 Cubic Clustering Criterion (CCC)

The cubic clustering criterion (CCC), a measure produced by the statistical software package, SAS, is the stopping rule implored in this research in combination with the pseudo F and $t^2$ statistics. 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 (9). The definition of $R^2$ is defined as:

$$\mathbf{R^2 = 1 - (P_G / T)}\text{, where}$$

- $\mathbf{P_G} = \Sigma W_j = \Sigma_{i \in Cj} \left| \mathbf{x}_i - \mathbf{x}_{ave(j)} \right|^2$, where summation is over G clusters at $G^{th}$ level
- $\mathbf{T} = \Sigma_i^n \left| \mathbf{x}_i - \mathbf{x}_{ave} \right|^2$
- $\left| \mathbf{x} \right|$ = Euclidean length of vector **x**, or the square root of the sum of squares of elements of **x**
- $\mathbf{x_i}$ = ith observation
- $\mathbf{x}_{ave(j)}$ = Mean vector for cluster $C_j$
- $\mathbf{x}_{ave}$ = Sample mean vector
- $C_j = j^{th}$ cluster
- G = Number of clusters at any level of hierarchy
- n = Number of observations

Based on the detailed comparative evaluation of stopping rules (14), Milligan and Cooper concluded that a ratio for between cluster variance to the within cluster variance provides a superior index for determining the optimum number of clusters. Sarle also provides

studies on the superior performance of the cubic clustering criterion (27). The CCC definitions below are taken from Sarle, where it is stated that the total sample variance along the $j^{th}$ dimension of the hyperbox is proportional to $s_j^2$ and the within cluster variance along the $j^{th}$ dimension is proportional to $c^2$. Sarle also states that "The CCC is based on the assumption that clusters obtained from a uniform distribution on a hyperbox are hypercubes of the same size. The hypercube assumption is obviously false in most cases, but is generally conservative unless the number of clusters is very large in two or more dimensions." (27)

$$\mathbf{CCC = \{\ln[(1 - E(R^2)) / (1 - R^2)\} * \{((nP/2)^{.5}) / ((.001 + E(R^2))^{1.2})\}, \text{ where}} \quad (27)$$

- $\mathbf{E(R^2)} = 1 - [(\Sigma^{p^*}_{j=1}(1/(n + u_j)) + (\Sigma^{p}_{j=p^*+1}(u^2_j / (n + u_j))) / (\Sigma^{p}_{j=1} u^2_j)] * [(n - q)^2 / n] * [1 + (4 / n)]$
- n = Number of observations
- q = Number of clusters
- p = Number of variables
- $s_j$ = Edge length of hyperbox along the $j^{th}$ dimension
- v = Volume of hyperbox
- $v = \Pi^{p}_{j=1} s_j$
- c = Volume of hyperbox
- $c = (v/q)^{1/p}$
- $u_j$ = Number of hypercubes along $j^{th}$ dimension of the hyperbox
- $u_j = s_j/c$
- $p^*$ = Dimensionality between clusters, $p^* < q$

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 TOD intervals to be developed based on historical traffic conditions.

### 3.8.2 Pseudo F and $t^2$ Statistics

The pseudo F and $t^2$ statistics are output in SAS when the data are coordinates or when using the centroid, average or Ward cluster method. The F statistic for a given level is calculated according to the following formula (10).

$$\textbf{Pseudo F} = (((\Sigma_i^n \, | \mathbf{x}_i - \mathbf{x}_{ave} |^2 - P_G / G - 1)) / (P_G / (n - G)) \text{ , where}$$

- $| \mathbf{x} |$ = Euclidean length of vector **x**, or the square root of the sum of squares of elements of **x**
- $P_G = \Sigma W_j$, where summation is over G clusters at $G^{th}$ level of hierarchy
- $W_j = \Sigma_{i \in Cj} \, | \mathbf{x}_i - \mathbf{x}_{ave(j)} |^2$
- $\mathbf{x_i}$ = ith observation
- $\mathbf{x}_{ave(j)}$ = Mean vector for cluster $C_j$
- $\mathbf{x}_{ave}$ = Sample mean vector
- $C_j = j^{th}$ cluster
- G = Number of clusters at any level of hierarchy
- n = Number of observations

This calculation takes into account the between and pooled within cluster sum of squares. The following formula shows the calculation for the pseudo $t^2$ statistic (10).

$$\mathbf{t^2} = B_{KL} / ((W_K + W_L) / (N_K + N_L - 2)), \text{ where}$$

- $B_{KL} = W_M - W_K - W_L$
- $N_K$ = Number of observations in $K^{th}$ cluster
- $N_L$ = Number of observations in $L^{th}$ cluster
- $W_k = \Sigma_{i \in Ck} \, | \mathbf{x}_i - \mathbf{x}_{ave(k)} |^2$
- $\mathbf{x_i} = i^{th}$ observation
- $\mathbf{x}_{ave(k)}$ = Mean vector for cluster $C_k$
- $C_k = k^{th}$ cluster

The pseudo $t^2$ statistic, which can be taken from Duda and Hart's Je(2) / Je(1) stopping rule, considers the sum of squared errors within clusters, the standard normal score, the number of dimensions and the sample size (9). It is important to note that these statistics

are not distributed as random variables since the cluster algorithms do not assign clusters randomly.

### 3.8.3 Recent Cluster Stopping Rule Studies

It has been stated, “There are no completely satisfactory methods for determining the number of population clusters for any type of cluster analysis (Everitt 1979, 1980; Hartigan 1985; Bock 1985), (15). Studies have been conducted to test the validity of clusters at different levels for determination of an appropriate hierarchical clustering level. These stopping rules have not produced a clear-cut solution to this problem due to the problem of ordinary significance tests failing with cluster testing. For instance normal ANOVA tests do not hold up under the assumptions imposed because cluster algorithms attempt to maximize the separation between clusters and the formation of clusters is not random (15). The stopping rules recommended in this project come from studies done in the 1980’s. This section discusses more recent stopping rule studies.

In the last decade, further research has been conducted to find the optimal method of determination of an appropriate hierarchical level. Atlas and Overall propose a method of evaluating higher-order cluster analyses in 1994, on cluster means from split-sample cluster analyses to determine the number of clusters using a replication criteria (14). Perfect replication at any particular hierarchical level is defined as a solution in which a single cluster mean from each of the preliminary analyses is grouped into each higher-order cluster. This method was compared to the Calinski and Harabasz pseudo F statistic, which performed the best in Milligan and Cooper’s comparison of 30 stopping rules. Atlas and Overall discovered that both methods uncovered the correct number of clusters for well separated populations; however, their study investigated the use of

overlapping clusters. Much of the work to evaluate stopping rules has involved an unrealistic separation between clusters. The following claim was made by Atlas and Overall: "At present, the perfect replication criterion applied to results from a higher-order clustering of means from several preliminary cluster analyses appears superior in its ability to determine the number of discriminably different underlying multivariate normal populations. Further evaluative work is perhaps needed, and we would hope that the replication criterion provided by higher-order cluster analysis can be included there as well (14)." Due to the investigative nature of this research, the replication criteria will not be investigated thoroughly in this project but will be recommended for future research. Sensitivity analyses in *Section 5.4* address the number of clusters determined by the pseudo F, T2 and CCC stopping rules. These rules uncover appropriate levels of the cluster hierarchy for the purpose of signal plan development and TOD intervals for volume and occupancy traffic data.

### *3.9 Cluster Analysis Input Data*

The preliminary data analysis was done on a small data set consisting of approximately 126 data points, with the final analysis being done with a data set on the order of 1000 observations to see how cluster formations are affected with different sample sizes. As the cluster analysis is performed on larger data sets, certain concerns must be considered. For example, clusters formed from a period of over 6 months, as is the case with the formal data analysis in this project, may produce clusters that contain observations over similar times of day in different clusters. This may be due to traffic variance over time and variant conditions occurring due to holidays and events and random days. Figure 19

shows a case where a large data set produces these repetitive clusters. Since the goal of this research is to base timing plans and TOD intervals on large historical data sets to capture a realistic picture of traffic conditions, it is important to remedy this situation.

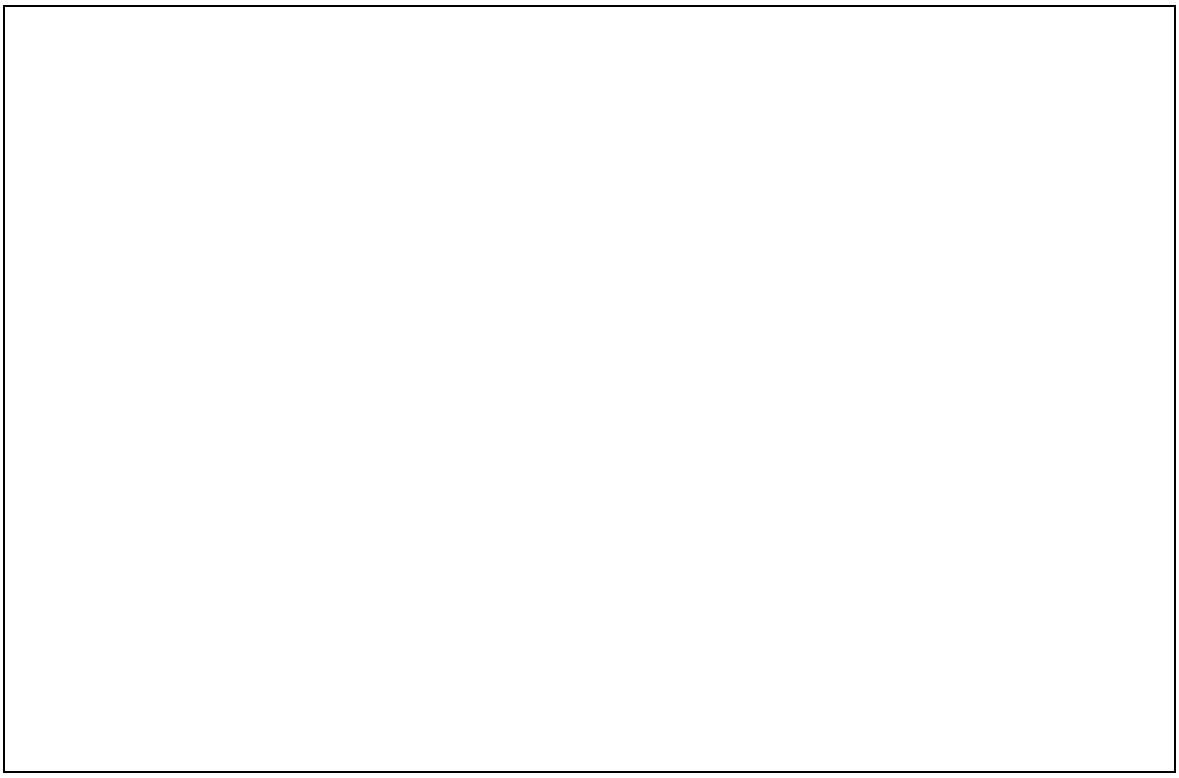


**Figure 19. TOD Intervals with Large Data Set**

One possible solution to the case described above is that different clusters can contain different densities of observations, thus promoting the use of one cluster for particular times of day over another even though they both may contain observations at the same TOD. For example, in Figure 19, cluster 1 and 2 appear nearly repetitive, but cluster 1 may contain 700 observations at 22:00, while cluster 2 may only contain 10 observations at that time. This would suggest the use of cluster 1 at that time; however, such an observation is impossible to make from the above graph. Thus, the suggested procedure will be to take the mean of all volumes and occupancies at each 15-minute

interval from the historical data set used for the cluster analysis prior to clustering. To do this, the assumption must be made that the volumes and occupancies are normally distributed about the mean at each 15-minute time interval. The following figures show an example of a volume distribution at an individual time period. Figure 20 and Figure 21 show the distribution of volumes at 7:15 over a 6-month period compared to a normal distribution. Figure 20 shows the volume distributions with a normal curve and Figure 21 shows the same volume distributions with the red bars showing how those volumes would be distributed for a normal distribution. The normal fit to these variable distributions according to 'Expert Fit' is 95% accurate and a "Good" fit, thus validating the averaged TOD method for cluster analysis (26). Table 1 displays the statistics associated with this volume distribution.

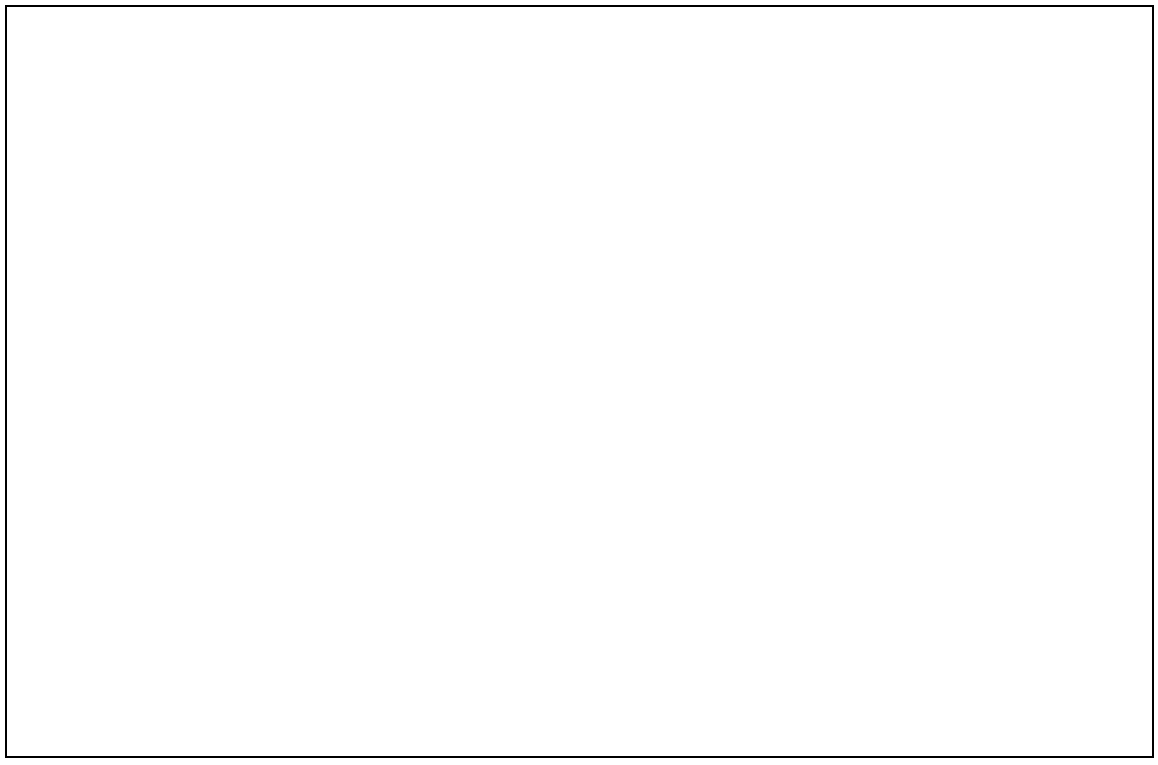


**Figure 20. Volume Distribution with Normal Curve at 7:15**

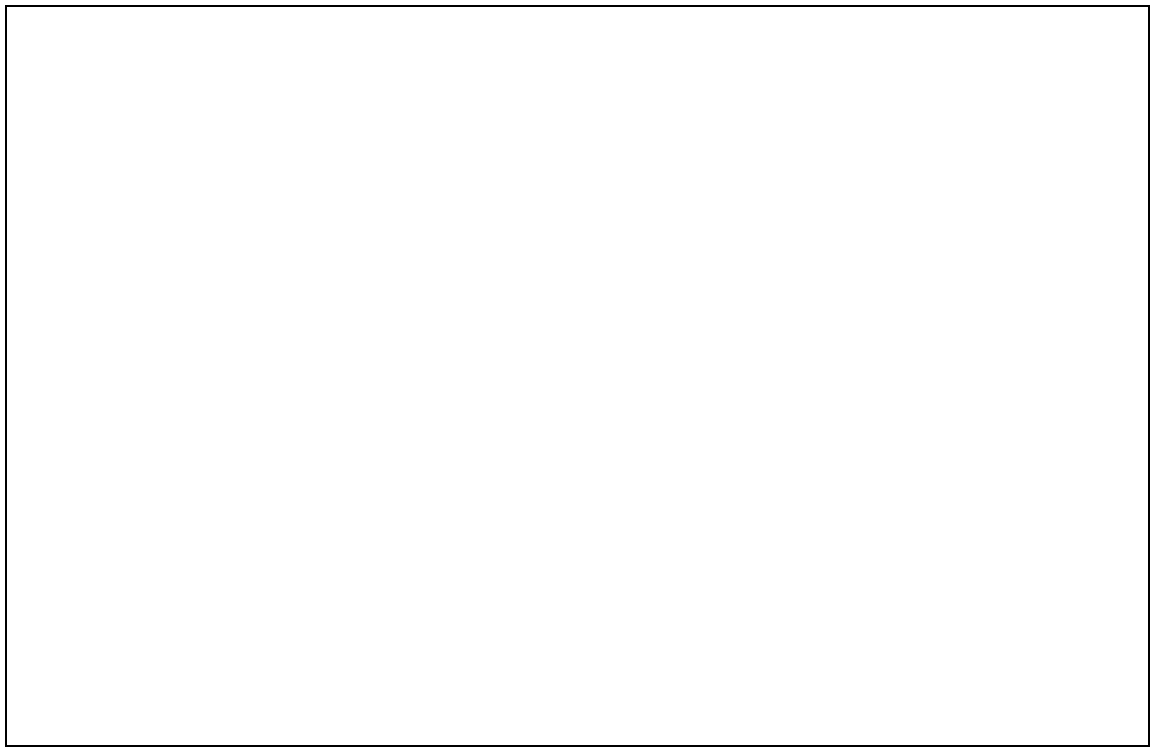
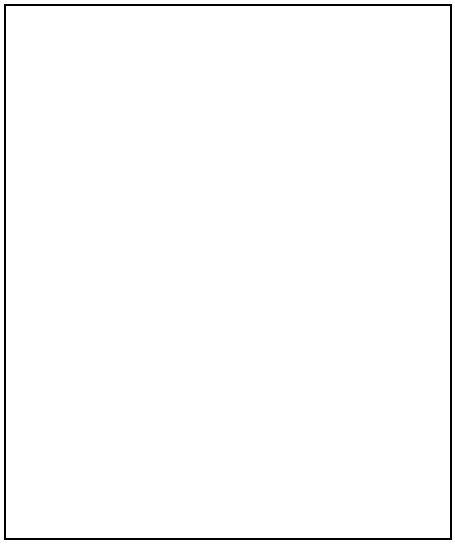


**Figure 21. Volume Distribution Compared to Normal Distribution at 7:15**

**Table 1. Descriptive Statistics for 7:15 Volume Distribution**



All of the data follow a normal distribution as shown in the example above. Thus, it is viable to use the mean of the observations in the data set at each 15-minute time

interval for cluster development to avoid situations of repetitive clusters as seen in Figure 19. The 95$^{th}$ confidence intervals for the averaged volumes from the three-intersection case study are displayed on the Volume vs. TOD plot in Figure 22. This shows that the range that 95% of the historical volumes lie in about the mean is fairly compact. The largest ranges occur during the peak periods of the day. Since the intervals lie fairly symmetrically about the mean, it is viable to assume an average value at each TOD is a good representation of traffic at that time. The plot of all volumes that exist during each TOD corresponding to this plot can be viewed in Figure 9.

The confidence intervals are fairly small, so the choice of timing plans may not be influenced by the occurrence of a maximum versus a minimum volume at most times. It is possible that during the peak periods, especially, the occurrence of a minimum volume may suggest the use of a pre/post-peak period plan; however, unless this was a very regular occurrence (which is highly unlikely), the suggested peak plan would suffice during the peak periods.

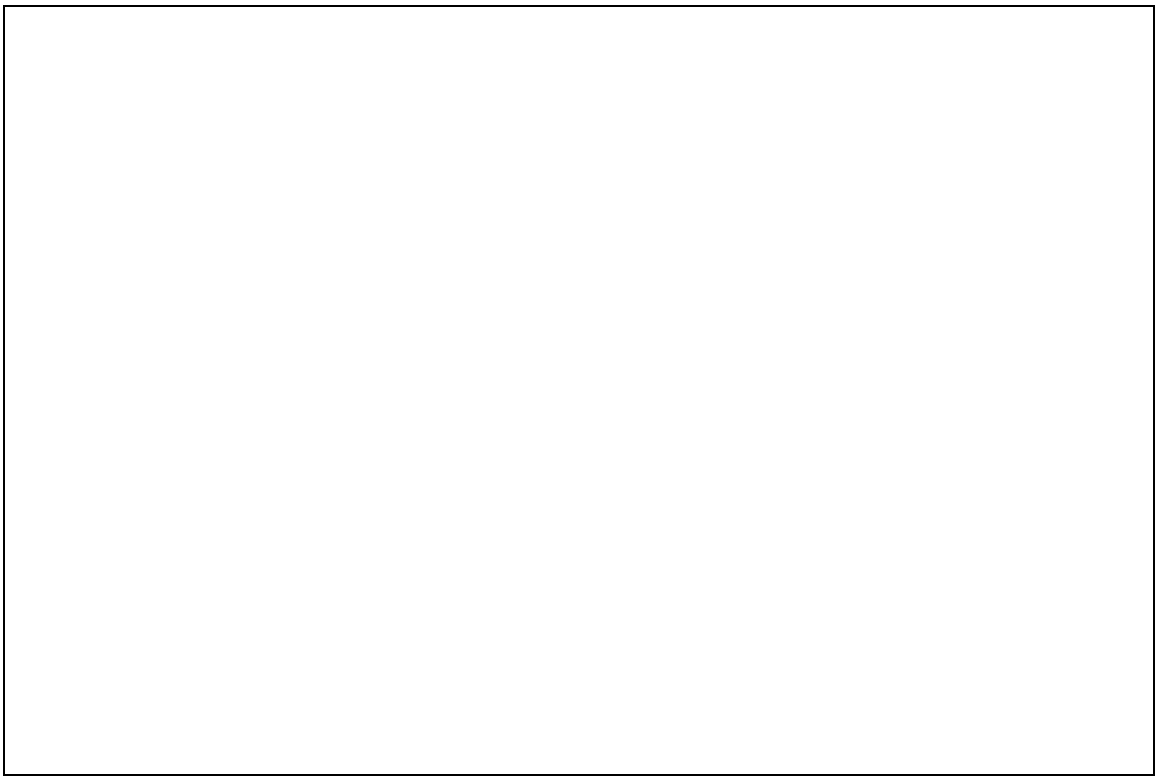


**Figure 22. NB Volume vs. TOD Plot with Confidence Intervals**

### *3.10 Cluster Validation*

It is extremely important to validate the cluster formations before drawing conclusions about the behavior of the data. Yet, cluster validation is commonly overlooked for several reasons. It is an extremely difficult problem due to the disagreement of what "cluster" means or what "validity" means (25). Since clustering is a tool used for discovery, rather than an end solution, it is common that inappropriate statistical models are chosen to validate the clusters. To make a substantial contribution to data analysis, it is essential that the clusters be validated to ensure meaningful conclusions.

Two methods for cluster validation have been implemented for evaluating the results of the cluster analysis on the 3-intersection case study. The first method is an internal validation criterion based on the raw data before and after cluster analysis and the

second is an external validation method through the use classification and simulation. An internal criterion assesses the fit between the structure and the data, using only the data (11). External methods may measure performance by rating a clustering structure using an outside tool such as simulation to assess the performance of the clusters under realistic conditions. The clusters formed by the 3-intersection case study are illustrated in Figure 25, from which TOD intervals are determined depending on when clusters are formed.

### 3.10.1 Internal Cluster Validation

Internal cluster validation should consist of two levels; validation by data and validation by imposed structure (25). The first task, validation by data, is to check the data for clustering tendency, or in other words, to ensure the data is not spatially random. This is important because clustering algorithms will produce clusters in any data set, whether it is completely random or contains some inherent groupings in the data. If a random data set gets clustered, the clusters would most likely be random themselves, holding little meaning. Projecting the variables onto a 2-dimensional space prior to clustering will show whether any natural groups exist in the data. The data set can also be broken into subsets and clustered. This should produce similar clusters if the data is not spatially random and the clusters formed are meaningful. Another measure for cluster tendency is the proximity matrix, which is a major data component used to validate clusters by the data themselves when testing for spatial randomness (25). The entries in a data matrix are indices of similarity, such as correlation, or dissimilarity, such as distance. The proximity matrix can be used to see if patterns exist in space prior to the clustering to support the assumption that the data is not random and should be clustered. The second task will judge the success of an algorithm in imposing a structure, assuming the data is

non-random. The following four structural criteria must be clearly defined for validation of the clustering structure imposed on the data (25).

1. Compactness: Measure the cohesion or uniqueness of an individual cluster.
2. Isolation: Measures the distinctiveness or separation between a cluster and its environment.
3. Global fit: Measures the accuracy with which the structure describes relationships between clusters, as well as the extent to which individual clusters are valid.
4. Intrinsic dimensionality: Determines the shape of a cluster and provides information about representing the patterns in a cluster.

A methodology for measuring these four criteria is not apparent, especially since the four criteria are not mutually exclusive but highly inter-related. A series of techniques will be presented to account for these criteria. Graphics for distance between opposing clusters and distance of observations within clusters will provide insight into the compactness and isolation of the cluster formations. The proximity matrix and the dendrogram will measure the global fit of the cluster analysis and the intrinsic dimensionality will be addressed by projecting the clustered data into a 2-dimensional space.

#### 3.10.1.1 Cluster Tendency

Testing for the tendency of the data to cluster can also be viewed as the test for "complete spatial randomness," a major piece of the validity tests. For this research, the principle components of the raw data were analyzed and the data projected onto the two primary principle components for a two-dimensional viewing of the cluster tendencies in the raw data. For the volume and occupancy data set from the three-intersection corridor case study, a principal component analysis was performed in SAS on the raw data and the projection can be viewed in Figure 23. The principal components or eigenvectors of the covariance matrix define a linear projection that replaces the features in the raw data with

uncorrelated features. The data can be projected onto the axes of the two largest eigenvalues, thus showing whether there is any natural grouping tendency in the data or not in a two-dimensional space (11). The eigenvalues represent the roots of the variance-covariance matrix. Due to the decreasing order of variance associated with the eigenvectors, it is typical that a summarization of the variability and covariability of the original variables from the two largest eigenvectors is sufficient (21). See Figure 23 for the 2-D projection of the raw volume and occupancy data from the three-intersection case study onto the axis of the primary principle components. From this figure the natural tendencies of similar times-of-day to group together are shown. See Table 2 for the corresponding times associated with each graph symbol. A series of similar symbols exist in the graph to represent an individual 15-minute interval from the time periods listed in the table. For example the number 2 represents 02:30 – 05:00, according to Table 2, and there are ten 2's in Figure 23. Each of the ten 2's represents a 15-minute time slice from the period of 02:30 – 05:00. The groupings of similar numbers shows the raw data exhibit a tendency to cluster during similar times of day based on the volume and occupancy pairs over a 24-hour period of the day.

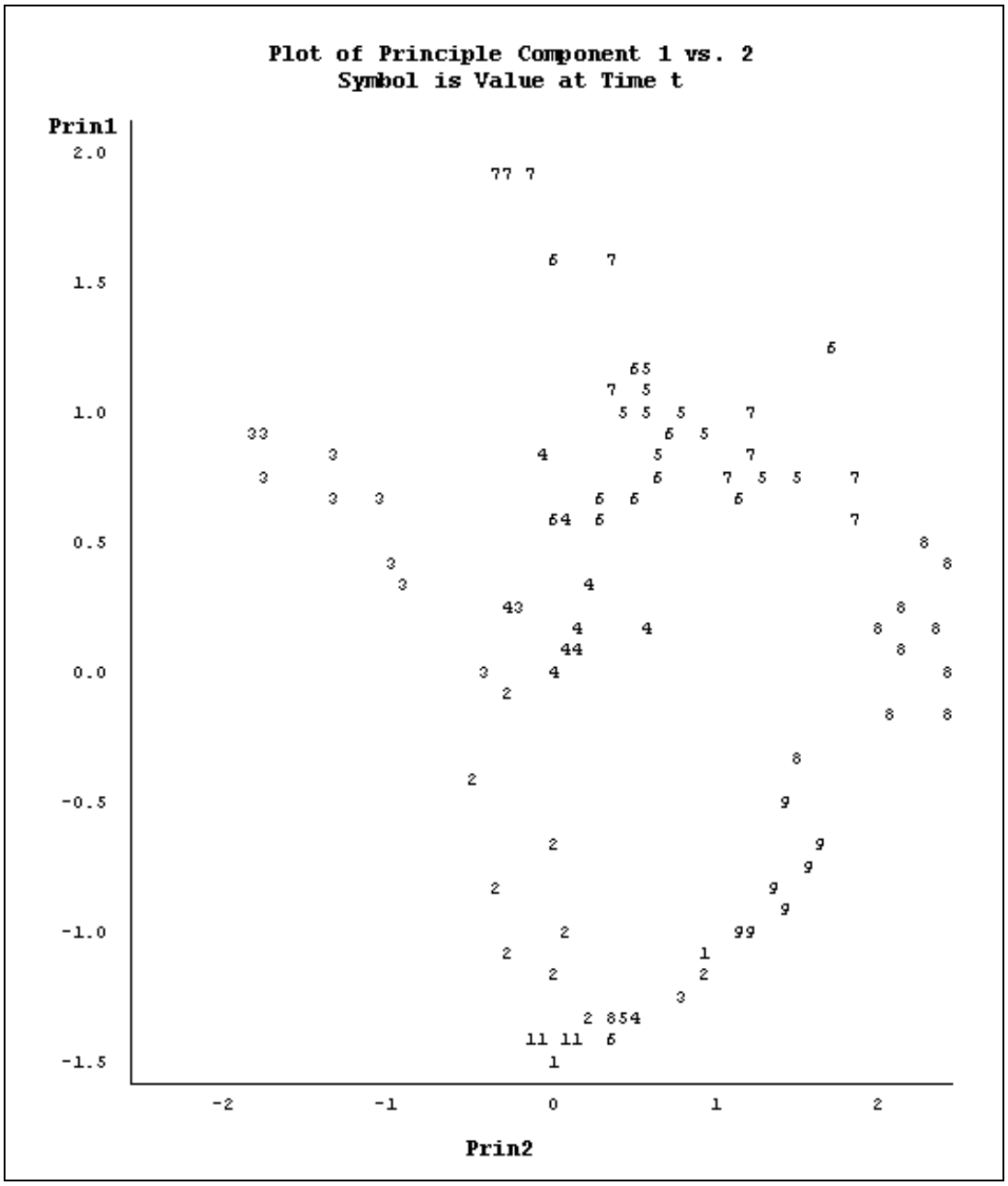



**Figure 23. Natural Raw Grouping Tendencies**

**Table 2. Graph Symbol Representations for TOD's from Figure 23**

Even before a data set is clustered, it should be determined if the data exhibit a "cluster tendency" or a predisposition to cluster into natural groups without identifying the groups themselves (11). This is an important consideration for cluster validity because clustering algorithms will create clusters whether data are naturally grouped or completely random, the latter resulting in meaningless clusters. Testing for cluster tendency essentially consists of testing the raw data for spatial randomness, which would infer the data is not appropriate for clustering. It is also possible for data to be regularly spaced, or to exhibit mutual repulsion, which would defeat the purpose of applying a clustering algorithm. The 3-intersection case study data set has shown to be non-random based on the natural groupings in the data over a 24-hour period as seen in Figure 23 and so the cluster validation process will continue to test for the validity of the cluster structure.

A method of establishing the stability of a cluster solution, another form assuring the data is not random, is to randomly divide the data set into sub-sets and performing a cluster analysis on each subset separately (22). Similar solutions should be obtained from both sets when the data is clearly structured. This technique was used successfully by Jolliffe, et al., 1982 (23). To demonstrate this approach, a sub-set of the original data set

from the three-intersection corridor was used for an example. The original 3-intersection data set consisted of volume and occupancy data from 8 March – 29 September 2000, while the subset data set consisted of data from 8 March – 1 July 2000. Figure 24 shows the TOD intervals formed from the cluster analysis for the full data set and Figure 25 shows the TOD intervals formed from the cluster analysis from the subset of that data set.

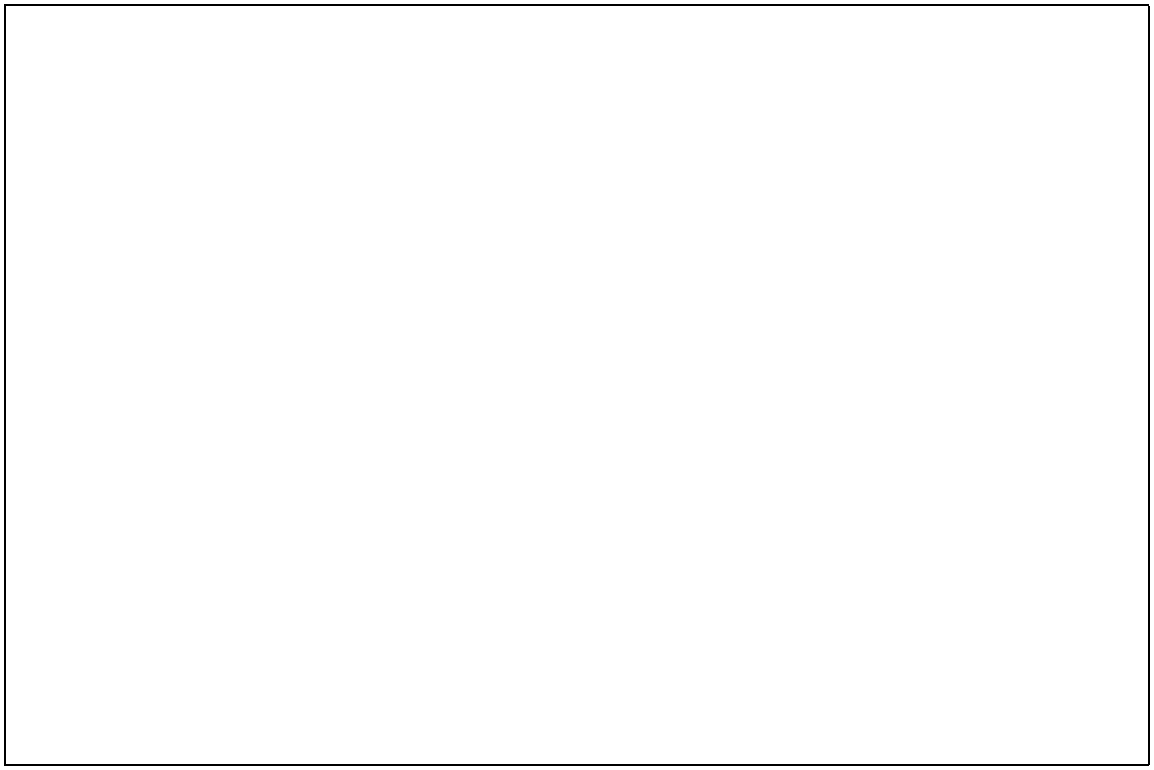


**Figure 24. TOD Intervals for Full, 3-Intersection Data Set**

**Figure 25. TOD Intervals for subset of 3-Intersection Data Set**

From the above charts, it is clear that both data sets created clusters distributed over nearly exact times of the day. Both clusters produced clear TOD intervals for the off-peak, AM, post-AM, Mid-Day, PM, post-PM, Evening and Pre-Off peak periods, where the mid-day and post-PM periods exist in the same cluster for each data set.

#### 3.10.1.2 Global Fit

Figure 26 shows the dendrogram produced for the 3-intersection cluster analysis. The dendrogram shows the level of dissimilarity along the y-axis at which point clusters join. The time-of-day associated with each initial cluster is displayed on the x-axis in a vertical, downward format. The times join in an intuitive manner as far as typical TOD traffic conditions exist. The dendrogram for the partial subset clustering is nearly the same as that for the full, 3-intersection data set. The dendrogram can also be viewed as an indicator for the degree to which a cluster formed is "real." A cluster can be termed

"real" if it forms early in the dendrogram for its size and lasts a relatively long time before being joined into another cluster (25). The smaller the dissimilarity (y-axis distance), the more alike observations in the cluster are. As the dissimilarity measure gets longer, the clusters are more likely to contain observations less similar. This is apparent in Figure 26 where the last levels of the cluster hierarchy join smaller clusters and the dissimilarity distance becomes elongated since clusters are being formed with all of the observations. This dendrogram was cut at the $7^{th}$ level according to stopping rules produced for the cluster analysis. The red line represents the $7^{th}$ level at which the dendrogram was cut. Determining if a cluster is "real" addresses the $3^{rd}$ point above about determining the global fit of the cluster. From the dendrogram, it appears that the clusters formed at the $7^{th}$ level consist of similar groupings of observations, with fairly small dissimilarity distances.

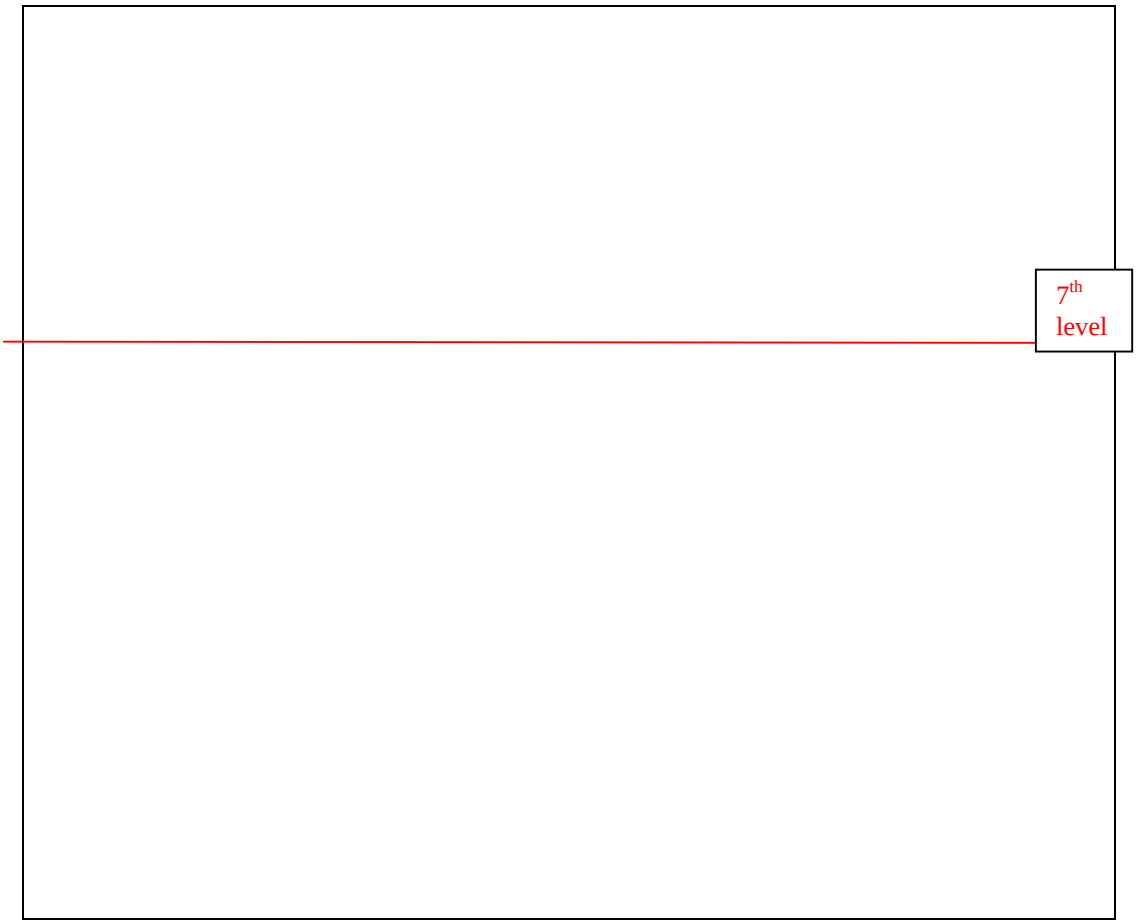



**Figure 26. Distance between Clusters for 3-Intersection Data Set**

### 3.10.1.3 Validation of Individual Clusters

The two main properties of clusters are compactness and isolation (11). A valid cluster would be one that was "unusually" compact and isolated, where compactness refers to the cohesion among objects within a cluster and isolation refers to the measure of separation between separate clusters. This section addresses the isolation criterion and the compactness criterion by visualizing these aspects of the resulting cluster analysis.

The first graphic for portraying the compactness and isolation of individual clusters is the distance of observations from their cluster centers and the distance of opposing cluster centroids from each other. Both of these aspects of individual cluster

validity are displayed in Figure 27. The first 7 characters in the legend represent the distance each observation in that cluster falls from its center. The remaining 21 characters in the legend represent the upper diagonal values of the distance between cluster centroids for all seven clusters to portray the between cluster distances. This display validates that the observations making up each cluster are much closer to their own cluster centers than opposing clusters are to other clusters. This supports the criterion that clusters must be compact and isolated. The clusters that fall closest to each other are intuitive. For instance, at cluster member 4, cluster 1 falls fairly close cluster 4, and these clusters represent the off peak and the pre-off peak periods. All of the closest cluster centroids follow this intuitive TOD scheme. The cluster member TOD classifications can be found in Table 3 for reference.

**Figure 27. Cluster Isolation and Compactness with Distance Measures**

Another visualization aid for assessing cluster solutions, suggested by Cohen (23), is to plot the cluster membership along the x-axis and above each cluster label, plot variable values for that cluster. Figure 28 shows the mean volumes for the cluster input variables that exist in each cluster. The error bars represent the standard deviations of each variable in each cluster. The variables are represented by 'V' or 'O' for volume or occupancy, a phase number and an intersection identifier, where 'SH' = Sunset Hill, 'BLMT' = Bluemont and 'ND' = New Dominion. Figure 29 shows the same plot, except with the occupancy means that exist in each cluster for each variable. The standard deviations are also represented in this chart as error bars on each variable.

Table 3 shows the times of day that each of the clusters represents. The variable values that exist in each cluster correspond to the TOD associated with each cluster. For instance, cluster 1 contains mean volumes and occupancies with the smallest values and this cluster represents the off peak period, while cluster 7, which contains the largest volume and occupancy values, represents the PM period. These two figures validate the criterion that the clusters must be unique as well as contain correlated variables. Each cluster is comprised of a combination of volume and occupancy values of which represent different movements at the different intersections. The volume and occupancy differences are obvious and intuitive for the timing plans they represent.

**Figure 28. Volume Means for 3-Intersection Clusters**

**Figure 29. Occupancy Means for 3-Intersection Clusters**

**Table 3. TOD Classifications for 3-Intersection Corridor**

Figure 28 and Figure 29 provide a visual for the cohesiveness of the variables within each cluster and the differentiation between opposing clusters. To present an even clearer portrayal of the separation between clusters, Figure 30 shows the overall mean and standard deviation of the volume and occupancy values present in each cluster. Again, the increasing volume and occupancy values with peak periods are apparent here. This plot is not as detailed since the individual movement volume and occupancy values that make up each cluster are averaged to one value.

**Figure 30. Cluster Mean Volumes vs. Occupancies for 3-Intersection Case**

The final display for cluster compactness is a demonstration of one of the cluster variable compositions for all 7 clusters. Figure 31 also shows the distribution of the variables within each cluster. The assumption can be made that the variables are normally distributed about the mean of each cluster centroid. This figure demonstrates this trend with an example from the volume variables at the Bluemont intersection, in the northbound direction. The cluster 2 and 5 volumes at this intersection, in the northbound direction, as well as clusters 3 and 6, contain similar, overlapping volume values. However there are many more variables than just this one contributing to the formation of these clusters, so this is not an issue.

**Figure 31. Variable Distribution within Clusters**

#### 3.10.1.4 Intrinsic Dimensionality & Isolation Criteria

The following figure follows the same idea as that of Figure 30 to represent the location of the cluster in a 2-dimensional space, only in a more detailed diagram. Each of the

volume and occupancy pairs for every movement in the 3-intersection corridor are displayed in their corresponding cluster in Figure 32. The projection was made on the two primary canonical variables, derived from the cluster membership values in SAS. The cluster groupings follow the same pattern as those in Figure 30, where similar TOD periods are located closest. The only difference in the following graph is that each observation making up each cluster is included in Figure 32. The numbers in this graph represent cluster membership values, whose TOD classification can be viewed in Table 3. This figure validates the isolation of the cluster formations as well as providing insight into the shape of the clusters formed.

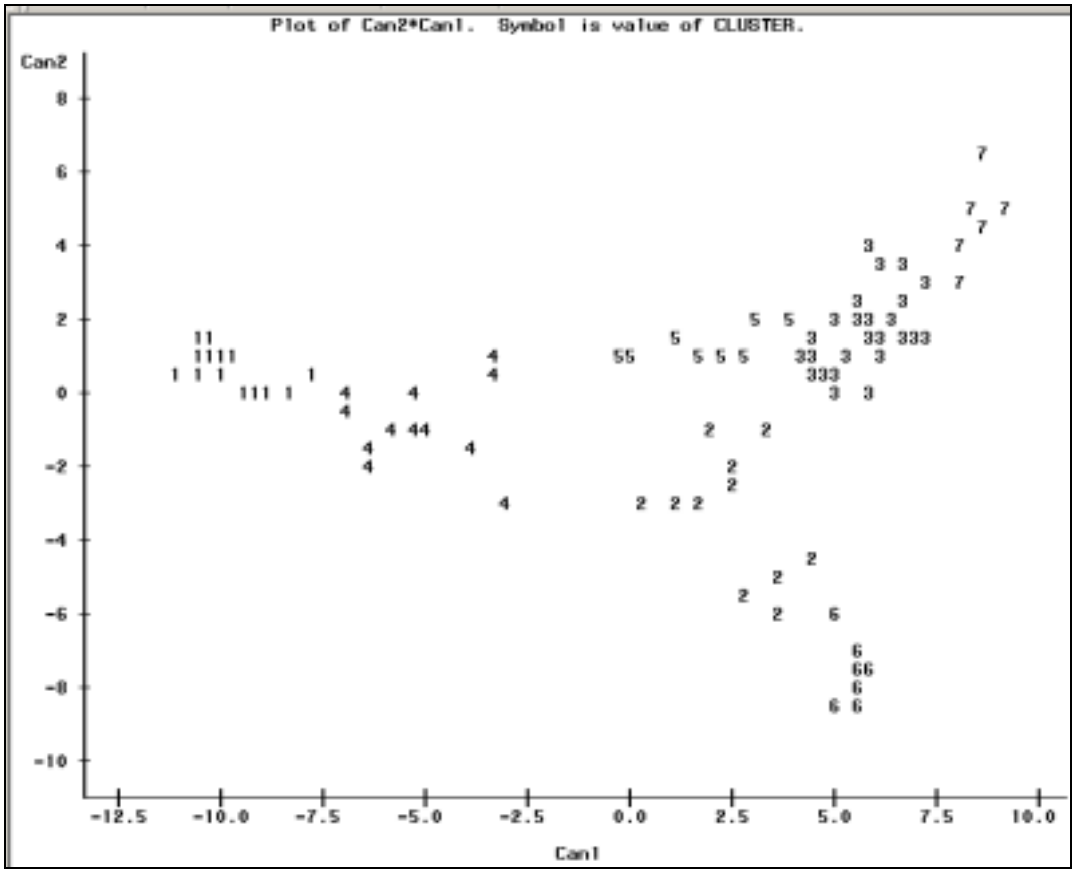



**Figure 32. 2-D Projection of Clustered Variables**

#### 3.10.1.5 Validation of Hierarchies (Global Fit)

The cophenetic correlation coefficient (CPCC) has been proposed for quantitative data for determining if the results of a hierarchical clustering are good, in particular to validate the hierarchies imposed by the clustering algorithm (11). The proximity between objects i and j can be called d(i, j), and the cophenetic proximity, to be called $d_c(i, j)$, is the level in the dendrogram for a particular clustering method where objects i and j are first placed in the same cluster. The CPCC is the product-moment correlation coefficient between the entries of these two matrices. These matrices are symmetric and so only the entries above the main diagonals are computed. The value of CPCC is between –1 and 1, and the closer it is to 1, the better the match between the two matrices and the better the hierarchy fits the data. Appendix A contains the matrices of CPCC values for each of the 7 clusters in the 3-intersection corridor analysis. It would be expected that if the hierarchy split at the $7^{th}$ level is a good cluster fit, then the CPCC values in the matrix should be close to 1, implying the variables in that cluster were joined at a level appropriate for the minimum distance of those clusters. It is important to recall that the variables making up the clusters represent opposing traffic movements and so certain movements should be less correlated than others should. For instance if the cluster to represent the AM peak is examined, some cluster variables represent northbound movements and some represent southbound movements. The AM peak northbound traffic will be much heavier due to the location, south of the business area, whereas the southbound AM traffic will be much smaller at that time. Of course the opposite becomes true during the PM peak period. Therefore, it would be expected that less correlation exist between opposing movements than like movements. This relationship

must be considered when examining the proximity and correlation matrices. The cophenetic correlation coefficient matrices in Appendix A follow these guidelines, in that the expected traffic movements with heavy flow at certain times, contain CPCC values close to 1. This implies the hierarchical cluster fit to the data is a good fit.

The validation methods investigated above provide the necessary insight into the clustering tendency of the data and the overall fit of the cluster hierarchy. By showing the stability of the cluster formations and the isolation and compactness of the clusters formed, it follows that the clusters formed are based on real grouping tendencies in the data. The groupings formed also follow a traffic condition intuition for the behavior of traffic during a 24-hour period during the week. The number of clusters that should be formed, or the level at which to cut the tree is another important issue that will be investigated in the 'Sensitivity Analysis' section in Chapter 5. Some external forms of cluster validation include testing the classifications formed by the cluster analysis, as well as simulating the formation of clusters as timing plans to test for actual performance of the groups formed. The next section deals with rating the cluster memberships formed with classification models.

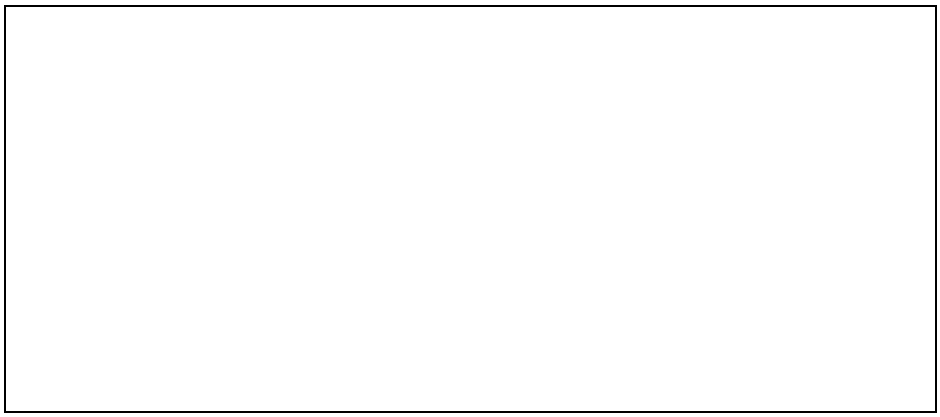
### 3.10.2 Secondary Cluster Validation – CART

Classification and Regression Trees (CART), version 3.6.3, is a classification tool that can be used as a secondary method of cluster validation. CART is based on decision tree technology that automatically searches for patterns and relationships and uncovers hidden structure in data. This information can then be used for predictive modeling, which is a useful and important piece for the future work of this project to be discussed in Chapter

6. Classification is appropriate since the target value, which will be introduced as the cluster membership developed in SAS, exists as a categorical variable representing a timing plan. Based on the input volume and occupancy variables at each detector associated with each cluster membership value for each observation, CART will construct a classification tree for classifying observations that can be used without cluster memberships' to a cluster or timing plan. Cross-validation is used for constructing the classification rule in CART due to the limited size of the data sets, which consist of only 96 observations (to cover a 24-hour period). The success of the classification rule on the cross-validated data will supply a secondary form of cluster validation since it can be assumed that a highly successful classification rule would imply meaningful clusters were formed (21). This secondary validation technique was performed on the two case study cluster outputs: the single intersection at Baron Cameron and Reston and the 3-intersection corridor.

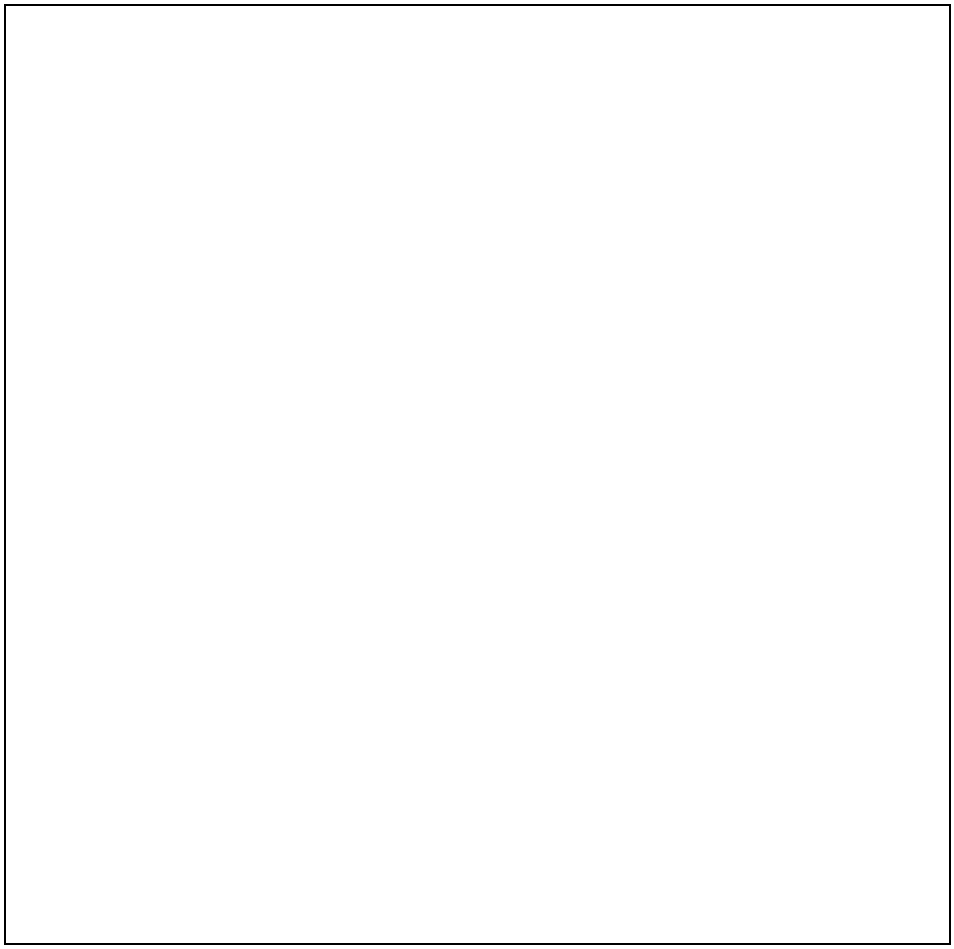
For both of the case studies, the classification results were superior with equal priors. The tree was selected based on the minimum cost of the tree and the GINI method was implored for splitting the data. In the single intersection case study at Baron Cameron, it was determined that four clusters were optimal, so this is the number of levels for the target variable in CART. Table 4 shows the cross-validation classification table for the Baron Cameron intersection. The total correctly predicted observations by the classification rule developed are 96.9%. Observations in cluster 1 and 4 were classified correctly 100% of the time, while the observations in clusters 2 and 3 were never classified less than 90%. This single intersection classification of the 96 clustered observations implies that the data were clustered into meaningful groups by SAS.

**Table 4. CART - Cluster Validation at Baron Cameron**



This secondary cluster validation technique with classification was also performed on the 3-intersection case study. This data set also contains 96 observations but this set was clustered with 7 clusters so the target variable for the classification consisted of 7 levels. Table 5 shows the cross-validated classification table for the 3-intersection classification. The overall percent of correctly classified observations with cross-validation was 92.7%. The clusters that performed the worst for correct classification are those with very few observations in them, thus producing low classification results. Only cluster 4 and 5 mis-classified more than one observation, with cluster 4 mis-classifying 3 observations and cluster 5 mis-classifying 2 observations. The overall classification rate for the example is good at 92.7% even with a small sample size and a fairly large number of clusters. This again supports the clusters produced in SAS as valid by secondary validation with classification.

**Table 5. Cluster Validation at Three-Intersections**



### 3.10.3 External Cluster Validation – Simulation

Clusters developed in the cluster analysis were also validated using simulation. Since the clusters represent TOD intervals, the performance of these newly created intervals under actual traffic conditions provides feedback as to the validity of the clusters formed. The aim is that the use of a refined state definition and accumulated historical data will develop more appropriate TOD intervals to be determined by the cluster analysis. Figure 33 shows the measures of performance from the 3-intersection simulation. The legend is explained below Figure 33 and can be referred to for the Chapter 4 analysis of results as well.

**Figure 33. SimTraffic Outputs for 3-Intersection Case Study**

- **Old Plan, Old TOD** – Plan developed with hand-counted volumes, implemented during the handpicked TOD intervals based on critical intersection traffic.
- **Old Plan, New TOD** – Plan developed with hand-counted volumes, implemented during newly clustered TOD intervals based on full state definition.
- **New Plan, New TOD** – Plan developed with database volumes from 6 months, implemented during newly clustered TOD intervals based on full state definition.
- **New Plan, Old TOD** – Plan developed with database volumes from 6 months, implemented during newly clustered the handpicked TOD intervals based on critical intersection traffic.

The TOD intervals represented in the legend as 'New TOD' were developed from the cluster analysis. The 90$^{th}$ percentile volumes for these centroids were used for Synchro timing plan development. These newly optimized plans based on cluster volume values are represented as 'New Plan' in the legend. 'Old TOD' represents the current TOD intervals used by VDOT and 'Old Plans' represent the current timing plans developed from the single-day, hand counts. It is clear that the new TOD's with the new plans

perform better than the Old TOD's with the New Plans, supporting the claim that the clustered TOD intervals perform better in simulation than those chosen be traffic engineers with the aggregate volumes at the critical intersection. When comparing the Old TOD's and Old Plans with the New TOD's and Old Plans, it is again shown that performance is improved with the newly clustered TOD intervals operating under similar timing plans with the old TOD intervals. The support of the simulation results for the improvement of performance with new TOD intervals can be viewed as a secondary, external validation that the clusters formed are logical and do provide a better form of defining TOD intervals based on data from all intersections in the corridor. These results will be discussed in more detail in Chapter 5.

### *3.11 Timing Plan Development and Simulation (Synchro/SimTraffic)*

Synchro/SimTraffic is a complete software package for modeling and optimizing traffic signal timing plans and then simulating these plans with the software, SimTraffic. Synchro and SimTraffic have been developed to provide simultaneous plan development and simulation. This software has been developed by Trafficware, (http://www.trafficware.com), a traffic signal software company. Synchro implements the methods of the Highway Capacity Manual to provide intersection capacity analysis and timing optimization where it optimizes cycle lengths, splits and offsets (16). This eliminates the need to try multiple timing plans in search of the optimum. Synchro optimizes to reduce delays and is the only signal software currently available that models actuated signals. Timing plans are developed in Synchro using historical data base volumes and output files for each timing plan are created for use in SimTraffic simulations to test the clustered plans that corresponds with each TOD interval.

SimTraffic is designed to model networks of signalized and un-signalized intersections (16). The primary purpose of SimTraffic is to check and fine tune traffic signal operations before implementing them in the field. SimTraffic includes the vehicle and driver performance characteristics developed by the Federal Highway Administration for use in traffic modeling (16). SimTraffic is especially useful for analyzing complex situations that are not easily modeled macroscopically including:

- Closely spaced intersections with blocking problems
- Closely spaced intersections with lane change problems
- The affects of signals on nearby un-signalized intersections and driveways
- The operation of intersections under heavy congestion

The following list summarizes the features modeled by SimTraffic (16):

- Pre-timed Signals
- Actuated Signals
- 2-way stop intersections
- All-way Stop intersections
- Freeways
- Roadway Bends
- Large Traffic Circles
- Lane additions and Lane Drops
- Cars, Trucks, Buses
- Pedestrians

SimTraffic is capable of simulating traffic conditions read in from outside files. These files are based on data base volumes at 15-minute intervals and the simulation effectively mimics the trend of traffic conditions according to these historical volumes. SimTraffic is also able to simulate transitions between timing plans by reading in the plan files output from Synchro that correspond to the times being simulated. The transitions occur according to the following steps (16):

1. New timing plan is loaded and cycle clock set based on time from midnight.

2. Cycle Clock for current state is calculated based on current phase durations and their start time.

3. The calculated cycle clock is compared to the target cycle clock. If the calculated cycle clock state is ahead by less than half a cycle, the controller will attempt to regain coordination by using shortened phases. Otherwise the controller will attempt to regain coordination by using longer phase times.

4. The transition max-times are calculated by increasing or decreasing the phase max green times by 17%. No green times will be shortened below the pedestrian walk plus flashing-don't-walk times or the minimum initial time. If shortening is unable to reduce the cycle length by at least 10%, the transition will occur using longer green times.

5. The signal will continue to time using the shorted or longer phase times. No force off or yield points are used.

6. At the beginning of each barrier transition, the calculated cycle clock is compared to the actual cycle clock. When the calculated cycle clock is a little bit behind, the transition is complete and the signal will begin operating coordinated with the new timing plan.

Signal transitions with pre-timed signals can be quite disruptive. It may take nearly an entire cycle to reach the sync point, then the signal may rest on the main street phases for up to a full cycle in addition to the normal main street green time. Thus, it is imperative that studies be conducted for network performance at transitional points, especially since this proposed procedure tends to portray an increased number of TOD intervals for which transitions must occur. A major draw back with SimTraffic is that only 19, 15-minute intervals can be simulated at one time, however there are no restrictions on the number of intersections in the network.

### 3.11.1 SimTraffic Outputs & Measures of Effectiveness

SimTraffic produces three main output elements for analysis of timing plan performance. The first is a performance report where delay, travel times, fuel emissions, etc. are reported. The second output is the queuing report, which includes the queuing

information at each movement. The final output is the signal report, which produces signal outputs from each phase in the system. Sample output files from each of the three reports are displayed in Figure 34, Figure 35 and Figure 36. The following is a list of Measures of Effectiveness SimTraffic provides in its reports:

- Slowing Delay
- Stopped Delay
- Stops
- Queue Lengths
- Speeds
- Travel Time and Distance
- Network Throughput
- Fuel consumption and efficiency
- Exhaust Emissions
- Observed Actuated Green Times

Each of these elements are included in the three main output files mentioned above and are discussed in full detail in the following sections.

#### 3.11.1.1 SimTraffic Performance Report

The performance report includes measures of performance for delay, stops, speeds, travel times, travel distances, number of vehicles and exhaust emissions. This is the main output report for use in this research for evaluating timing plan effectiveness. Figure 34 shows an example performance report created by SimTraffic.

SimTraffic Performance Report

Baseline 08/28/1899

Main Street & SB Ramp Performance by movement

| | EBT | EBR | WBL | WBT | SBL | SBT | SBR |
|---|---|---|---|---|---|---|---|
| Total Delay (hr) | 2.7 | 0.0 | 0.0 | 0.0 | 0.6 | 0.0 | 0.0 |
| Delay / Veh (s) | 127.5 | 10.4 | 3.3 | 3.2 | 20.0 | 25.4 | 3.8 |
| Stop Delay (hr) | 2.5 | 0.0 | 0.0 | 0.0 | 0.5 | 0.0 | 0.0 |
| St Del/Veh (s) | 117.8 | 6.8 | 0.3 | 1.7 | 17.0 | 19.8 | 2.8 |
| Total Stops | 116 | 5 | 1 | 2 | 72 | 3 | 10 |
| Stops/Veh | 1.55 | 1.25 | 0.10 | 0.06 | 0.69 | 1.00 | 0.67 |
| Travel Dist (mi) | 7.8 | 0.4 | 0.8 | 3.0 | 10.1 | 0.3 | 1.5 |
| Travel Time (hr) | 2.9 | 0.0 | 0.0 | 0.1 | 1.0 | 0.0 | 0.1 |
| Avg Speed (mph) | 3 | 14 | 21 | 23 | 10 | 9 | 19 |
| Fuel Used (gal) | 2.0 | 0.1 | 0.1 | 0.7 | 1.2 | 0.0 | 0.4 |
| Fuel Eff. (mpg) | 4.0 | 7.3 | 9.1 | 4.1 | 8.6 | 8.3 | 4.0 |
| HC Emissions (g) | 6 | 0 | 0 | 3 | 4 | 0 | 1 |
| CO Emissions (g) | 149 | 12 | 16 | 193 | 135 | 4 | 22 |
| NOx Emissions (g) | 15 | 1 | 1 | 9 | 11 | 0 | 2 |
| Vehicles Entered | 77 | 4 | 10 | 36 | 107 | 3 | 16 |
| Vehicles Exited | 73 | 4 | 10 | 37 | 102 | 3 | 15 |
| Hourly Exit Rate | 438 | 24 | 60 | 222 | 612 | 18 | 90 |
| Denied Entry Before | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Denied Entry After | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Main Street & SB Ramp Intersection Performance

| | EB | WB | SB |
|---|---|---|---|
| Total Delay (hr) | 2.7 | 0.0 | 0.6 |
| Delay / Veh (s) | 121.6 | 3.2 | 18.0 |
| Stop Delay (hr) | 2.6 | 0.0 | 0.5 |
| St Del/Veh (s) | 112.1 | 1.4 | 15.2 |
| Total Stops | 121 | 3 | 86 |
| Stops/Veh | 1.53 | 0.06 | 0.69 |
| Travel Dist (mi) | 9.3 | 3.8 | 11.9 |
| Travel Time (hr) | 2.9 | 0.2 | 1.1 |
| Avg Speed (mph) | 3 | 22 | 11 |
| Fuel Used (gal) | 2.0 | 0.8 | 1.6 |
| Fuel Eff. (mpg) | 4.1 | 4.6 | 7.5 |
| HC Emissions (g) | 7 | 3 | 4 |
| CO Emissions (g) | 161 | 199 | 161 |
| NOx Emissions (g) | 16 | 11 | 13 |
| Vehicles Entered | 81 | 46 | 126 |
| Vehicles Exited | 77 | 47 | 120 |
| Hourly Exit Rate | 462 | 282 | 720 |
| Denied Entry Before | 0 | 0 | 0 |
| Denied Entry After | 0 | 0 | 0 |

**Figure 34. SimTraffic Performance Report**

Total Delay is equal to the travel time minus the time it would take the vehicle with no other vehicles or traffic control devices (16). For each time slice of animation the incremental delay is determined with the following formula:

$$TD = dT * (spdmax - spd) / spdmax, \text{ where}$$

TD = Total Delay for time slice
dT = time slice = 0.1s
spdmax = maximum speed of vehicle
spd = actual speed

The maximum speed may be less than the link speed if a vehicle is within a turn, approaching a turn, or accelerating out of a turn. Total delay also includes all time spent

by denied entry vehicles while they are waiting to enter the network. Delay per Vehicle is calculated by dividing the total delay by the Number of Vehicles.

The Number of Vehicles is not a fixed number because some vehicles are in the area analyzed before the interval begins and some are in the area after the end of the analyzed period after the interval ends. Part of these vehicles delay is counted in prior and subsequent intervals and thus it is not fair to count these vehicles in the vehicle count for this interval. The Number of Vehicles is thus equal to:

$$nVeh = nX - 0.5* nS + 0.5 * nE, \text{ where}$$

nVeh = Number of Vehicles
nX = Vehicles Exited this interval
nS = Vehicles in area at start of interval
nE = Vehicles in area at end of interval

Per vehicle values for a network or arterial will be higher than their intersection components. If, for example, all vehicles are delayed at 3 intersections for 5 seconds each, the network delay per vehicle will be 15s.

The Stopped Delay is the sum of all time slices where the vehicles are stopped or traveling at less than 10 ft/s (3 m/s). Normally the Stopped Delay will be less than the total delay. Stopped delay also includes all time spent by denied entry vehicles while they are waiting to enter the network. Stop Delay/Vehicles is calculated by dividing Stop Delay by the Number of Vehicles.

The Total Stops is a count of vehicle stops. Whenever a vehicle's speed drops below 10 ft/s (3 m/s) a stop is added. A vehicle is considered going again when its speed reaches 15 ft/s (4.5 m/s). Stops /Vehicles is calculated by dividing the number of Stops by the Number of Vehicles.

The Travel Distance is simply a summation of the vehicle distance traveled. This distance includes the curve distance within intersections.

The Travel Time is a total of the time each vehicle was present in this area. The travel time includes time spent by vehicles Denied Entry.

The Average Speed is calculated by dividing Total Distance by Total Time. Average Speed is weighted by volume and includes stopped time and denied entry time. The time use in calculation for Average Speed does not include time spent by denied entry vehicles while they are waiting to enter the network. Average speed may thus be higher than Total Time divided by Total Distance.

Fuel Used is calculated with the fuel consumption tables. The fuel used in each time slice is determined by the vehicle's fleet (car, truck, or bus), speed, and acceleration. The Fuel Efficiency is calculated by dividing the Total Distance by the Fuel Used. Emissions data are calculated with the vehicle emission tables. The vehicle's speed and acceleration determine the emissions created in each time slice. The vehicles queued in the denied entry number are not accounted for in the fuel used calculation and so this value would be disproportionally smaller than MOP's such as travel time. There is no emission tables available for trucks and busses. SimTraffic assumes trucks and busses emit exhaust at three times the rate of cars.

Vehicles Entered and Vehicles Exited is a count of how many vehicles entered and exited the link or area in the interval(s). If this is a network or arterial summary, the Vehicles Entered and Vehicles Exited do not count a vehicle moving from one intersection to the next within the arterial or network. The Entered and Exited counts for a network or arterial will thus be less than the sum of the counts from each intersection.

The Hourly Exit Rate is the Vehicles exited at an hourly rate. If the intersection is above capacity and the input volume is not constrained upstream, this value might be used as the capacity for this movement

Denied Entry is a count of vehicles that are unable to enter a link due to congestion. The report lists the number of vehicles denied entry at the start and end of each period. Thus, to determine the number of vehicles denied entry during each time interval, the number denied entry before must be subtracted from the number denied after the interval. This is useful to see if congestion is getting worse or better. Denied Entry can also be used to determine the Network Throughput. In a congested network, lower values of Denied Entry indicate increased throughput and vice versa. This is a good determining factor for the effectiveness of timing plans. The higher the number of denied vehicles typically infers that those timing plans are performing worse.

#### 3.11.1.2 SimTraffic Queuing Report

The queuing report includes information on queues and blockages encountered by the vehicles in the simulation. The Queuing and Blocking report gives information about the maximum queue length for each lane and the percentage of time critical points are blocked. Figure 35 shows an example queuing report file output by SimTraffic.

Queuing and Blocking Report

Baseline 08/30/1999

Intersection: Main Street & SB Ramp

| Movement | EB | EB | EB | WB | SB | SB | SB |
|---|---|---|---|---|---|---|---|
| Directions Served | T | T | R | T | L | LT | R |
| Maximum Queue (ft) | 310 | 336 | 20 | 27 | 244 | 261 | 47 |
| Link Distance (ft) | 531 | 531 | | 362 | 499 | 499 | |
| Upstream Blk Time (%) | | | | | | | |
| Queuing Penalty (veh) | | | | | | | |
| Storage Bay Dist (ft) | | | 150 | | | | 150 |
| Storage Blk Time (%) | | 0.40 | | | | 0.01 | |
| Queuing Penalty (veh) | | 17 | | | | 1 | |

Intersection: Main Street & NB Ramp

| Movement | EB | EB | EB | WB | WB | WB | NB | NB | NB |
|---|---|---|---|---|---|---|---|---|---|
| Directions Served | L | T | T | T | T | R | L | LT | R |
| Maximum Queue (ft) | 88 | 52 | 51 | 99 | 75 | 38 | 69 | 47 | 109 |
| Link Distance (ft) | | 362 | 362 | 460 | 460 | | 330 | 330 | |
| Upstream Blk Time (%) | | | | | | | | | |
| Queuing Penalty (veh) | | | | | | | | | |
| Storage Bay Dist (ft) | 150 | | | | | 150 | | | 150 |
| Storage Blk Time (%) | | | | | | | | | |
| Queuing Penalty (veh) | | | | | | | | | |

Network wide Queuing Penalty: 18

**Figure 35. SimTraffic Queuing Report**

Queues are reported individually for each lane, no summing or averaging is performed between lanes. A vehicle is considered queued whenever it is traveling at less than 10 ft/s (3 m/s). A vehicle will only become “queued” when it is either at the stop bar or behind another queued vehicle. The Maximum Queue is the maximum back of queue observed for the entire analysis interval. This is a simple maximum, no averaging is performed. The maximum queue is calculated independently for each lane. The queue reported is the maximum queue for each individual lane, NOT the sum of all lanes’ queues. SimTraffic records the maximum back of queue observed for every two-minute period. The Average Queue is average of all the 2-minute maximum queues. Vehicles can stop when queued and when waiting for a mandatory lane change. SimTraffic tries to determine whether the stopping is due to queuing or lane changes. In some cases stopping for lane changes will be counted as queuing. Sometimes in SimTraffic and in real life, the lane changes and queuing behavior are closely interconnected.

The Link Distance is the internal distance of the link from stop-bar to stop-bar. This value will be less than the link distance defined in Synchro because it is the internal distance after subtracting the widths of the intersections.

Upstream Block Time is the proportion of time that the upstream end of the lane is blocked. There is a hot spot 20ft (6 m) long placed at the top of the lane. Every time slice that this hot-spot is occupied by a queued vehicle counts towards the block time.

The Queuing Penalty is a rough measure of how many vehicles are affected by the blocking. The Queuing Penalty is equal to the estimated volume of the lane times the percent of time the lane is blocked. The Queuing Penalty for a storage bay blockage is based on the volume of the adjacent lane. If a through lane is blocking a storage bay, the penalty is based on the volume of turning traffic. The Queuing Penalty is a quick way to quantify the affects of queuing. It can be used to show, for example, that Timing Plan A has less blocking problems than Timing Plan B. Queuing Penalty is not calculated for external links.

Storage Block Time is the proportion of time that a lane is queued at the top of the storage. There is a hot spot 20ft (6 m) long placed at the top of the storage bay. Through lanes adjacent to storage bays are also tracked. Queuing in the through lane can block access to the storage bay. Every time slice that this hot-spot is occupied by a queued vehicle counts towards the block time.

#### 3.11.1.3 SimTraffic Actuated-Signals, Observed Splits Report

The actuated signal report displays information about the actual times observed in actuated signals (16). This report can be used to show how an actuated signal will

perform with detailed modeling. This report can be helpful to compare the affects of adjusting gap settings, detector layouts, recalls and so on. Figure 36 shows an example report.

Actuated Signals, Observed Splits

Baseline 08/30/1999

Intersection: 1: Main Street & SB Ramp

| Phase | 1 | 2 | 4 | 5 | 6 | 8 | 12 | 16 |
|---|---|---|---|---|---|---|---|---|
| Movement(s) Served | WBTL | EBWB | SBTL | EBTL | EBWB | NBTL | SBTL | NBTL |
| Maximum Green (s) | 28.0 | 16.0 | 16.0 | 28.0 | 16.0 | 16.0 | 4.0 | 4.0 |
| Minimum Green (s) | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 |
| Recall | Min | None | None | Min | None | None | Min | Min |
| Avg. Green (s) | 28.0 | 16.0 | 16.0 | 28.0 | 16.0 | 16.0 | 4.0 | 4.0 |
| g/C Ratio | 0.35 | 0.20 | 0.20 | 0.35 | 0.20 | 0.20 | 0.05 | 0.05 |
| Cycles Skipped (%) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Cycles @ Minimum (%) | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 100 |
| Cycles Maxed Out (%) | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| Cycles with Peds (%) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Average Cycle Length (s): 80.0
Number of Complete Cycles : 6

**Figure 36. SimTraffic Actuated Signals, Observed Splits Report**

Each column in the figure above represents one signal phase. Movements are the lane groups served by this phase. Maximum Green is the maximum green time before this phase will max out and the green time will be given to the next phase. For a coordinated signal this is the maximum time before the signal will yield or be forced off. Minimums Green is the minimum green time that a phase must retain the green, even if no vehicles are detected. In Synchro this is called the minimum initial time. Recall is the recall for the phase. This will be Coord for coordinated, Max for Max recall, Ped for Pedestrian recall, Min for minimum recall, or None for no recall. Avg Green is the average of all green times. Skipped phases do not count. Green periods that begin or end in another interval do not count. g/C Ratio is the observed green time to cycle length ratio. Since there may be green time measured from cycles that fall partially outside this interval, an adjustment is used. The formula for g/C is as follows:

**g/C = TotalGreen / TotalCycles * NumCycles / (NumGreens+NumSkips)**

Cycles Skipped (%) is the percentage of cycles skipped by this phase. Green periods or permissive periods that begin or end in another interval do not count. Cycles @ Minimum (%) is the percentage of cycles that show for their minimum time. Normally these phases have gapped out. Green periods that begin or end in another interval do not count. Cycles Maxed Out (%) is the percentage of cycles that max out. This value also includes all cycles for coordinated phases and phases with Max Recall. Green periods that begin or end in another interval do not count. Cycles with Peds (%) is the percentage of cycles with a pedestrian call. If this phase has Pedestrian Recall all phases will have pedestrians. Green periods that begin or end in another interval do not count. Average Cycle Length(s) is an average of the cycle lengths modeled. For a coordinated signal, this is the actual cycle length. Number of Complete Cycles(s) is a count of the number of complete cycles modeled. Partial cycles do not count, although phases from partial cycles may count for individual phase statistics.

The outputs produced by SimTraffic can be used to analyze entire network performance over entire TOD intervals, network performance at specific TOD intervals, intersection performance at TOD intervals and phase movement performance at TOD intervals. The capability for analysis is in-depth and for the scope of this study will focus on the main elements in the performance report for comparing TOD interval and plan effectiveness of the entire corridor.

### *3.12 Chapter Summary*

This chapter provides the background and selection of methods for the proposed procedure. The clustering algorithms are presented with the support for the selection of the centroid methodology for cluster analysis. Typical cluster outputs are presented with emphasis on issues faced such as the production of "bad" clusters by the cluster algorithm. Stopping rules are introduced and the selection of the CCC, Pseudo F and Pseudo $t^2$ statistics are supported and detailed for selection of the appropriate number of clusters. This chapter also summarizes some cluster validation techniques and presents the results of these validations for the 3-intersection case study. The final piece of this chapter is a detailed report of the simulation tool implemented in this research and the outputs available from SimTraffic.

# Chapter 4. Proposed Procedure

Based on the research done with the Northern Virginia system detector data, a procedure has been proposed for developing improved timing plans through the use of data mining tools. Since no such procedures have been discovered for utilizing detector data being collected by many DOT's, a proposal has been developed for using this resource on the most widely implemented method of timing plan development, TOD. The procedure section outlines the tools used in this research for the procedure and the steps taken through each stage of the process.

Figure 37 depicts the proposed procedure in a flow chart format. This procedure can be refined and improved with extended research, but is at this time introduces a method of utilizing detector data to improve on existing methods of signal timing development.

## *4.1 Tools*

The tools to be used for the proposed procedure are as follows:

- Data Extractor Tool
  - Developed by the Smart Travel Laboratory at the University of Virginia (2000)
  - http://smarttravellab.virginia.edu/Data%20Extractor/home.htm
- Microsoft Excel
- SAS, Version 8
  - Developed by SAS Institute, Inc.
- Synchro, Version 4
  - Developed by TrafficWare
  - http://www.trafficware.com
- SimTraffic, Version 4
  - Developed by TrafficWare
  - http://www.trafficware.com
- Classification and Regression Trees (CART)

### *4.2 Proposed Procedure Flow Chart*

***[Print out separate page for this figure in Landscape]***

**Figure 37. Proposed Procedure Flow Chart**

### *4.3 Data Collection*

The use of the Data Extractor Tool will query the Oracle database housed at the University of Virginia in the Smart Travel Laboratory where 15-minute Volume, Occupancy and Speed data can be retrieved from all intersections in the Northern Virginia arterial network. The Data Extractor outputs the data to an Excel spreadsheet, which can be used for any necessary data manipulation and for storing the data sets. The procedure for using the Data Extractor is as follows:

- Select 'Nova Detector Info' from menu.
- Select 'intersection' or 'corridor' to collect data from necessary detectors.
    - The Data Extractor lists all detectors in each section and corridor with the phase movement of each detector.
- Add the detectors or intersections of interest to the extraction list for data gathering.
- Select the 'Extract Data' view from the menu.
- Set the date and time interval for the dates of historical data collection.
- Select weekdays, weekends or particular days for tailored data collection, for the case presented here select weekdays.
- Select 'continuous' or 'segmented' interval for continuous 24-hour sampling or a subset of the 24-hour period to be sampled over dates chosen, for this case select 'continuous'.
- Select all screening procedures for fully screened data. (See Help menu of Data Extractor or Chapter 1 of this report for details of data screening procedures).
- Select average volumes for each phase movement to obtain an individual value for each movement, thus eliminating the need to research the number of lanes that exist in each movement since detectors may not exist in all lanes.
- Select "Get Data" and select "Data Formatting – Graph Format" to return data in a usable format in Excel.
- Save Excel output to file.

- In the Excel file, insert an addition column by the Datex column for time-of-day. The 'convert to text' option can be used to separate the date and time in the Datex column to only times-of-day. This is necessary for proper TOD identification after the cluster analysis.

- Average similar TOD variables to represent one mean value at each TOD for cluster analysis.

### *4.4 SAS Procedure for Cluster Analysis*

The statistical software package, SAS is used for producing clusters from the data to represent timing plans. The timing plans will exist for the time-of-day intervals to be specified by the cluster analysis. The hierarchical cluster procedure will be used for cluster development.

- Import the saved Excel data file into SAS. The import data function is in the File menu.

- Use 'Procedure Cluster' in SAS to develop clusters from the data file. A sample code for use in SAS can be viewed in Table 6.
    - The 'Standardize' option should be used to standardize all variables to mean = 0 and a standard deviation of 1 prior to the clustering process since volume and occupancy variables lie on a different scale.
    - The 'NPRINT = 9' option can be used to display only the statistics for the final 9 clusters since, due to hardware constraints, a maximum of 9 timing plans can be produced.
    - The 'CCC' option is used to display stopping statistics for use in determining the number of clusters to produce.
    - The 'PSEUDO' option can be used to display a pseudo F and $t^2$ statistic for aiding in the determination of the optimal number of clusters to produce.
    - The volume and occupancy variables at each movement at each intersection are the input variables and should be read in as they exist in the data file.
    - The 'ID = TOD' option should be used to copy the time-of-day values associated with each cluster assignment for identification of TOD intervals.

**Table 6. SAS Cluster Procedure (Code Example)**

- View the CCC, Pseudo F and $t^2$ statistics from the output file from the Cluster Procedure to determine the proper number of clusters to form. The cluster level with the first local maxima Pseudo F statistic, the largest CCC value and a small Pseudo $t^2$ value should be the optimal number of clusters to form.
    - The number of clusters to form can be determined with an expert rule. This rule should be based on the fact that the CCC provides an accurate descriptor of the appropriate number of clusters, with its inaccuracy exhibiting too many clusters as the appropriate number. The Pseudo F and $t^2$ statistics are also accurate with the mis-classifications occurring with too few clusters identified as the appropriate number. So if the maximum CCC, Pseudo F and minimum Pseudo $t^2$ do not occur at the same level, these factors can be accounted for in the expert rule for an automated selection of the number of clusters.

- With the proper number of clusters chosen, run the Tree Procedure in SAS. A sample of ProcTree can be viewed in Table 7.
    - The 'Dock = n' option should be used to require a minimum number of observations to exist for cluster formations, thus reducing the creation of clusters with too few observations. The n variable should be chosen according to the sample size of the data set, for this research an n value of 4 is used.
    - The 'Method = Centroid' cluster method should be used for the cluster analysis (The Ward method is also a good choice and produces very similar results).
    - The 'Nclusters = n' option should be used, where n = the appropriate number of clusters to be formed as decided from the output statistics of the Cluster Procedure (CCC, Pseudo F and $t^2$).
    - The 'Copy TOD, *input variables*' option should be used to copy the TOD's and volume and occupancy values associated with each observation and cluster.

**Table 7. SAS Tree Procedure Cluster Code**

- When the clusters have been formed, the Means Procedure should be run on those clusters to determine the descriptive statistics associated with each cluster. See Table 8 for a sample SAS code for the ProcMeans procedure.

- In the File Menu, Export the data tables produced with the Tree Procedure and the Means Procedure to an Excel Spreadsheet.

**Table 8. SAS Mean Procedure Cluster Code**

### *4.5 Determination of TOD Intervals*

The clusters produced in SAS will be used to determine the TOD intervals for improved timing plan development. The Excel file that was output from the Tree Procedure in SAS should be used to identify the TOD intervals.

- Format the 'TOD' column to time-of-day (hh:mm).

- Graph the times-of-day on the x-axis and the cluster membership on the y-axis to produce a graph of the TOD intervals as determined from the cluster analysis.

- The cluster analysis will produce patterns in the graph where the TOD intervals exist and these transitions can be used to represent the new timing plans.

### *4.6 Synchro Timing Plan Development*

The timing plans for use with the newly developed TOD plans via the historical data will be developed in Synchro. Volumes for each movement in each intersection for the corridor under development must be determined. These volumes must be those that exist for the TOD intervals developed in the cluster analysis. These volumes should also service the densest portions of those TOD intervals. Thus the 90$^{th}$ percentile volume values from the data set are used. The output data file from the 'Tree Procedure' in SAS can be used to determine these values existing for each cluster or timing plan. Once the timing plans have been developed and optimized for each of the TOD intervals, the timing files can be written to an Excel spreadsheet for use in SimTraffic. These files include split, cycle length, offset and lead information. This will allow the simulation to account for transitioning between plans during the 24-hour period. The Synchro files for Northern Virginia have been obtained from VDOT, such that all lane geometry's and statistics, plan statistics, driver characteristics, vehicle types, etc. have been developed accurately for proper timing plan development. The following list details the timing plan development procedure.

- A timing plan will be developed in Synchro for each cluster developed.
- The volumes for each timing plan are obtained from the volume values that make up the observations contained within the cluster that represents the timing plan being developed.
- Organize the SAS ProcTree output data set such that the data is sorted by cluster membership.
- Use the 'Percentile = (Volume data, .9)' to search for the 90$^{th}$ percentile of the volume data making up the cluster or timing plan being developed.

- Find the 90th percentile volume for each movement at each intersection.

- This 90th percentile value will ensure that the timing plan developed will accommodate the heaviest traffic conditions during that time period. The Maximum volume is not used to ensure that an erroneous case is not used.

- For intersections where detectors do not exist at every movement, a turning conversion obtained from current VDOT Synchro files must be used.

- These conversion factors should eventually be validated with data collection counts from the 'CAMVAN.'

- The Northern Virginia Synchro files obtained from VDOT are being used, while only altering the input volumes that the plans must accommodate. The existing volumes in the Synchro files were obtained from one-day physical hand counts. Through-movement to turning-movement conversions for all intersections in the Reston corridor have been developed from the existing volumes in the Synchro files. These conversion factors are used to infer turning movement volumes where detectors do not exist from the through-movement lanes, where detectors always exist for the newly developed timing plans inputs. Another method is ratio of change from known detector data over TOD's for each movement.

- To input these new volumes at an intersection click on that intersection and then hit the timing window icon and the lane configurations and volumes will be visible.
  - Use this value in the Synchro file, multiply by the number of lanes in each movement and multiply by 4, to represent the entire flow (VPH) for each movement since the data being used for clustering is an average value for each movement based on 15-minute intervals.

- On the left side of the screen under "options", make sure that the type of controller is actuated, coordinated and make sure that the "lock timings" box is not selected.

- Once the new volumes have been input, select the "optimize" tool bar from the top menu and optimize the splits, offsets and cycle length, for each intersection.

- Optimize network cycle length and offsets for entire network.

  - On the left hand side of the screen the timing plan characteristics such as cycle length, V/C ratio, intersection delay, etc. can be seen.

- With the new timing plans in place for the clustered TOD intervals, the timing files can be written to and Excel spreadsheet by selecting 'Data Options' from the 'Transfer' tool bar. Make sure the 'Timing' sheet is selected and select the location for the timing file, which should be written in (.csv), comma delimited format. Assign the timing plan a name in the 'Timing Name' section for use in SimTraffic. Hit the 'Write' button and the timing plan file will be available for use in SimTraffic.

    - Before running SimTraffic, make sure all TOD interval timing plans are placed in a single Excel spreadsheet.
    - Check the timing files to ensure the maximum splits match those in Synchro as well as the cycle length and offset information.

- With the new timing plans in place, the 'Animate' button can be selected from the icon menu bar and the plans will be used for simulation in SimTraffic.

## *4.7 Validation of Timing Plans with SimTraffic*

SimTraffic will be used to simulate the newly developed timing plans for the newly developed TOD intervals. Average 15-minute volumes from each time interval during the 24-hour period obtained from the SAS ProcTree output data set will be used to develop the clusters to feed into the simulation for setting up the parameters for the proper number of vehicles. These 15-minute volumes must be determined for each movement at each intersection to account for changing traffic patterns during each TOD interval.

### 4.7.1 Preparing 15-minute data tables for simulation

- Determine time interval that the specific plan is implemented.

- Organize the SAS ProcTree output data file by cluster and then by time-of-day.

- Find the average volume for each 15-minute interval in each timing plan at each movement.

- If detector data does not exist for all movements at the intersection, then use the conversion factors from the Synchro files to produce the missing volumes. Or, use original Synchro volumes for turning movements at peak TOD for each plan period, fluctuating the value over time to match the fluctuation of the detector volumes, available.
    - Transfer these values to an Excel spreadsheet.
    - The intersection ID can be obtained from Synchro by highlighting the intersection and selecting the "#" icon from the toolbar.
    - The date column can identify the date for which the data was collected.

- The excel file must be saved as a .csv file (comma delimited) for use with SimTraffic.

### 4.7.2 Preparing SimTraffic Parameters

- After pressing the “Animate” icon in Synchro, you will be transferred to SimTraffic.
- Stop the simulation and select the options menu and then select “Database Access” to prepare for inputting the 15-minute volumes.
- Go to “Data Options” and select “read Volumes from UTDF file” and find the location of the 15-minute volume table that was created for simulation. Make sure the data format option for .csv file is selected.
- Select the date that appears in the 15-minute volume file.
- In the “Data Options” sheet, also select the “read Times from UTDF file” and locate the timing file that was written from Synchro.
- Go to the intervals tab and make sure seeding is set at 0 for random seeding or select specific numbers for the seeding when performing multiple runs.
- Insert enough intervals to represent the correct length of time that the simulation will be run for.
- Change the duration time to 15 minutes if that is the length of volume intervals being read in from the file.
- Make sure that the times correspond to those times on the excel file.
- Allow an initialization of the simulation by seeding for at least three minutes without recording, prior to the interval start time, to allow the system to fill.
- Select the appropriate timing plan ID’s as written from the Synchro file to coordinate with the times being read from the volume file.
- Press the animate simulation icon and once the simulation is complete, go to file and create a report. A text version of the report can be saved.
- MOP’s such as travel time, number of stops, total delay, fuel used and travel distance are output by SimTraffic for the entire time period as well as for each 15-minute interval. These outputs are available for the entire corridor or for each individual intersection and movement. Any level of analysis can be performed using the outputs.

### *4.8 Development of Classification Rule using CART (Future Research)*

Once new timing plans have been developed based on historical data, a classification rule can be developed for classifying future cases into a pre-determined timing plan. Classification and Regression Trees (CART) is the tool used to create the classification rule by imploring binary splits on the data. The cluster membership value developed with cluster analysis is used as the response variable, while the volume, occupancy pairs are used as the input variables in the CART model. Further research should be conducted in this area to verify classification rules developed will handle the classification of future traffic states. Guidelines should also be established for number of mis-classifications necessary for out-dated plans and/or the need to adjust TOD intervals.

- Import data from cluster output.
- Use cluster membership as target variable.
- Use volume and occupancy variables as predictor variables.
- Use cross-validation for tree development.
- Use equal priors.

### *4.9 Chapter Summary*

The procedure detailed in this chapter directs the user through every step of the signal plan development procedure. Figure 37 summarizes this chapter into a flow chart that can be followed for developing timing plans with system detector data and data mining tools. The first three levels of Figure 37 will be automated during the summer of 2001 in the Smart Travel Lab to be delivered to the northern Virginia signal control group. This will allow the TOD intervals to be determined via cluster analysis and historical detector data with the push of a button. This will also produce the $90^{th}$ percentile volumes associated with each cluster formation for plan development in Synchro. At this stage of

development of the automated procedure, the tool will produce formatted, Excel files for importing into Synchro and Simtraffic with all of the lane turning movement volumes and timing plan TOD intervals necessary to produce plans and simulate them based on the detector data. The use of classification for determination of out-dated timing plans will be a separate study conducted in the Smart Travel Lab, with the automation of this process existing as part of the NOVA map, which is under development in the Smart Travel Lab. This is expected for completion in the summer of 2001.

# Chapter 5. RESULTS AND ANALYSIS

## *5.1 Introduction*

This chapter will investigate alternate input variables for the cluster analysis and the effects of the input variables on the cluster outputs in the *'Sensitivity Analyses – Cluster Input Variables'* section. These sensitivity analyses will suggest the form of cluster input variables that produce the cleanest TOD clusters. Sensitivity analyses will also be conducted for the investigation of the 'minimum number of observations' constraint imposed in the SAS cluster analysis. This section will look at the effects of imposing such a constraint on the cluster analysis and suggest an appropriate value for this constraint. The final sensitivity analysis conducted will investigate the selection of the appropriate number of clusters based on the stopping rules implored in this research. The levels at which to cut the cluster tree will be evaluated to verify that the stopping rules do in fact suggest the appropriate number of clusters for formation. The analysis of the exploratory case studies for a single intersection and a 3-intersection corridor will then be presented. The 3-intersection corridor will be presented with full detail of the significance of the results, while the single intersection case study will be presented for suggesting the usefulness of this procedure at single intersections versus corridors to be discussed in the *Conclusions* section. The single intersection case study will not provide the level of detail of analysis that the corridor supplies due to its less significant role in this project.

### *5.2 Sensitivity Analyses – Cluster Input Variables*

One of the capabilities of cluster analysis for the development of TOD intervals and timing plans is that it can utilize all available data in creating timing plans for new TOD intervals. This data includes volume and occupancies, which are available at lanes that contain system detectors. Some of the sensitivity analyses done here include creating clusters using different input variables. Standardized volumes and occupancies were initially used as cluster input variables. However, the cluster analysis may produce better clusters without occupancy or with it weighted less than volume, since occupancy really only provides useful information from values of 0% – 25%. Values greater than 25% may skew the resulting clusters. Occupancy greater than 25% means the roadway is saturated. For example, it is common for occupancy of 25% to have the same meaning as occupancy of 90%, whereas occupancy of 5% represents quite different conditions than occupancy of 20%. The cluster analysis was done using the initial volume and occupancy variables, which were standardized and thus equally weighted. A comparison of results was done using only standardized volumes, standardized volumes and occupancies, standardized volumes with occupancies converted to values < 26 to create clusters. These three cases were then compared with the variables un-standardized to utilize the natural weighting of the volumes and occupancies inherent in their data representation of traffic conditions. Finally, a case was done using weighted volumes, where the volume and occupancies were first standardized. All cluster analyses were constrained to only producing clusters containing at least four observations and the cluster methodology imposed was the centroid method.

The clustered TOD intervals produced in this section contain some gaps in time during the 24-hour period from which one plan transitions to the next (See Table 9). This

is due to the situation in which too few cases comprise a cluster or a small time slice is represented by a particular cluster. For the proposed procedure and automation of the plan development procedure, expert rules will be introduced to account for such situations. Small clusters that cannot support the development or transition to a timing plan will be assigned to the cluster occurring immediately before and after such an occurrence.

### 5.2.1 Standardized Input Variable Cluster Analyses

Figure 38 shows a cluster analysis done with standardized volumes and occupancies at one intersection in the Reston Corridor. Table 9 shows the TOD classifications for this cluster analysis. Five clusters were formed, with a constraint imposed on the clusters of a minimum of four observations existing in order for a cluster to be formed. Clusters 1 – 4 are intuitive as far as peak periods go. The fifth cluster does not make sense without looking at the data that makes up cluster 5. The reason for its formation is that the phase 6 (northbound) occupancies in cluster 5 are all greater than 50%. All other occupancies never get much higher than around 20%. So, the formation of cluster 5 is useless for developing a timing plan for that particular timing period due to the randomness of the times associated with the observations making up cluster 5. In reality, cluster 5 should probably be part of cluster 4, or a PM period, but the occupancies seem to confuse the clustering process here.

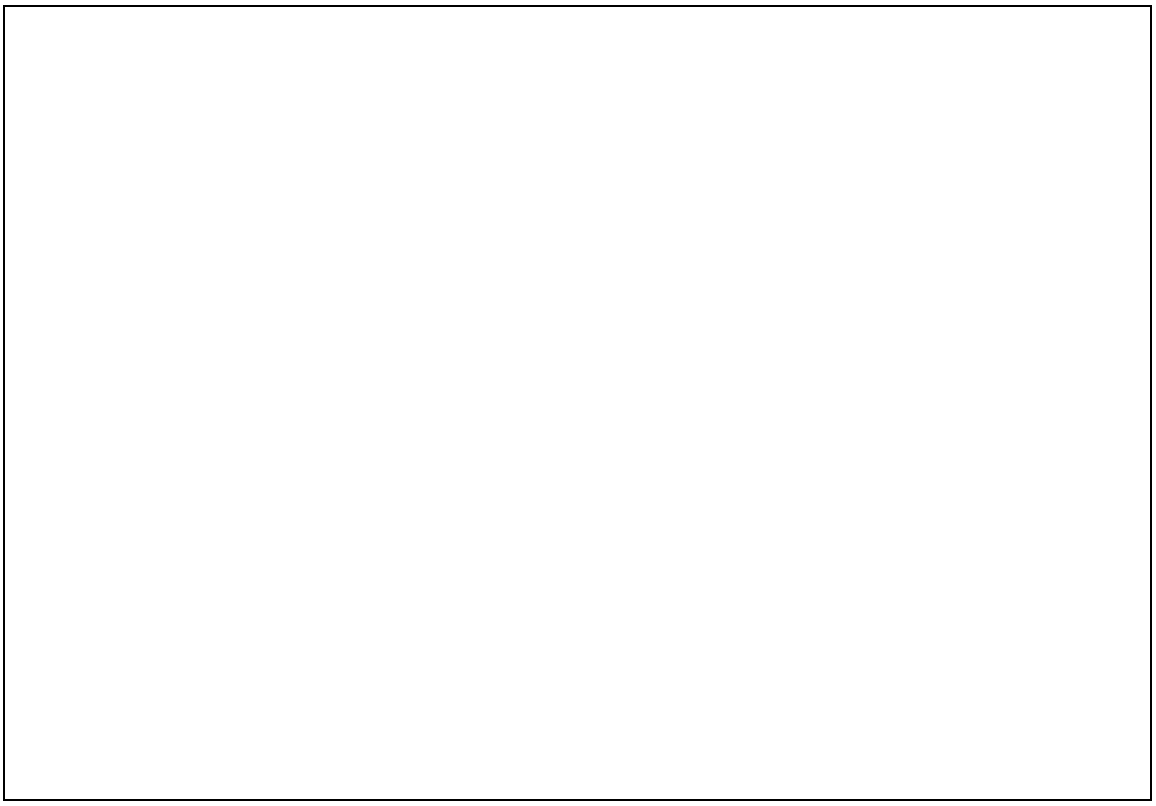


**Figure 38. Clustering with Standardized Volume & Occupancy**

**Table 9. TOD Classification for Volume & Occupancy Cluster**

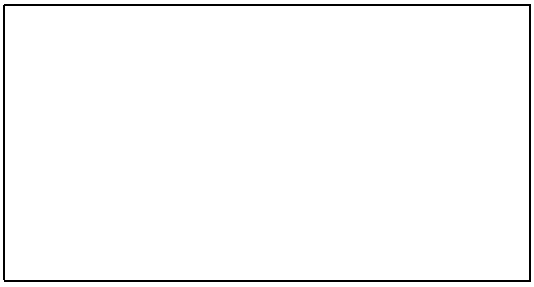


Figure 39 shows a cluster analysis done on the same data set where only the standardized volumes were used as input variables. The TOD classifications can be viewed in Table 10. The erroneous looking cluster 5 that was formed in Figure 38 does not appear in Figure 39 since occupancies are not included. Cluster 5 in the following cluster analysis is representative of a PM peak period that was not captured in the cluster

analysis in which volume and occupancy was used. The first four clusters in both analyses capture similar TOD intervals and transitions between timing plans. Erroneous data hurts the cluster analysis and it may not be a good idea to use the occupancies in the cluster analysis unless the data is further cleansed before clustering. For instance, all occupancies greater than 25% could be assigned a value of 25% to represent saturation, thus eliminating situations as that in Figure 38.

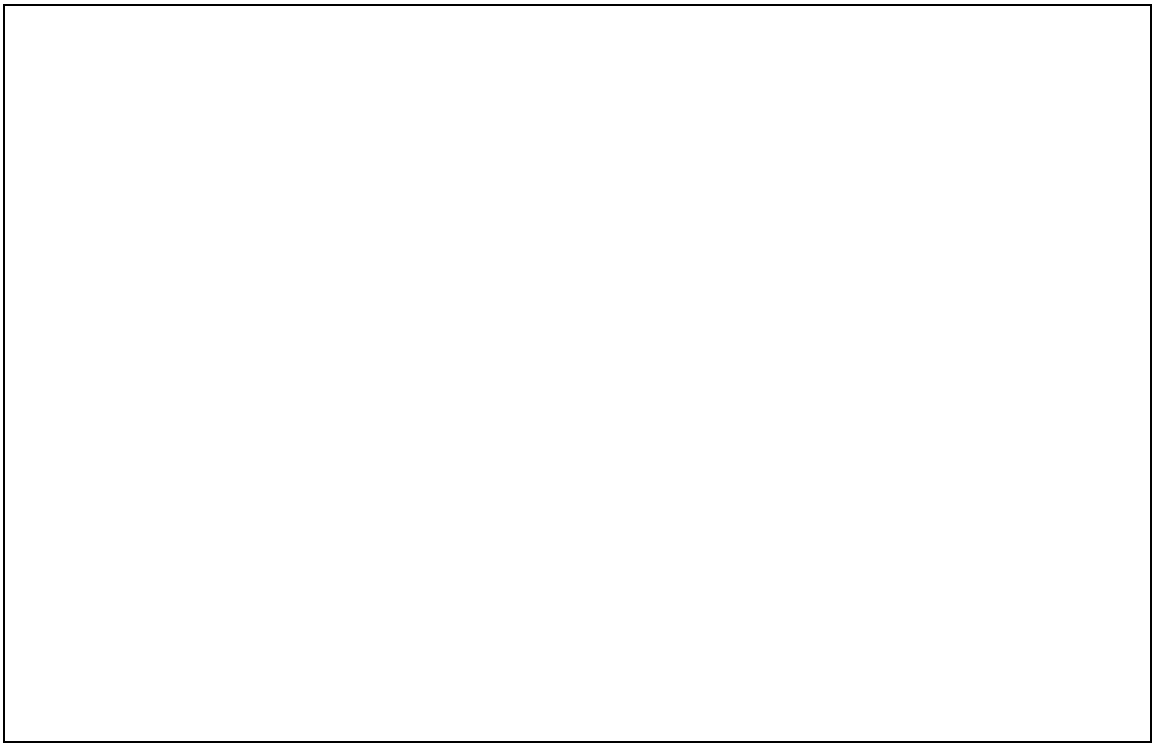


**Figure 39. Cluster Analysis with Standardized Volumes**

**Table 10. TOD Classification for Volume Cluster**

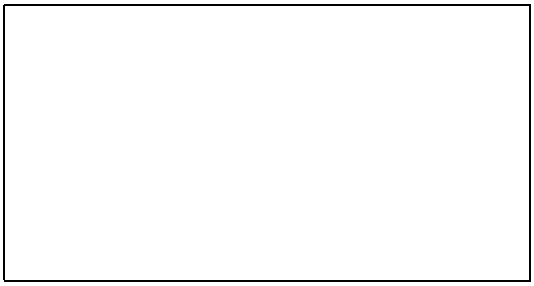


Figure 40 shows a cluster analysis with the standardized volume and occupancy values; however, the occupancies have been adjusted such that all occupancies < 26. As discussed above, since any occupancy of approximately 25% or greater represents saturation, all saturated values were converged to 25%. This eliminates the erroneous cluster 5, assigning it PM peak values as in Figure 39. The TOD intervals and timing plans can be classified as follows in Table 11.

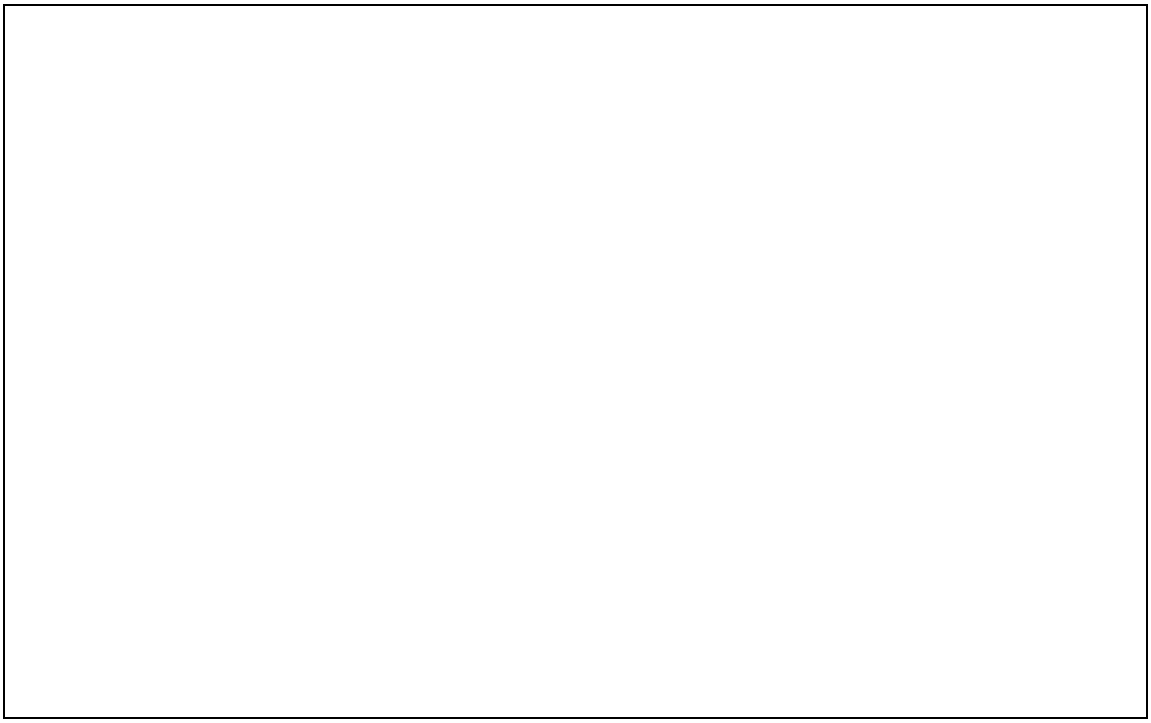
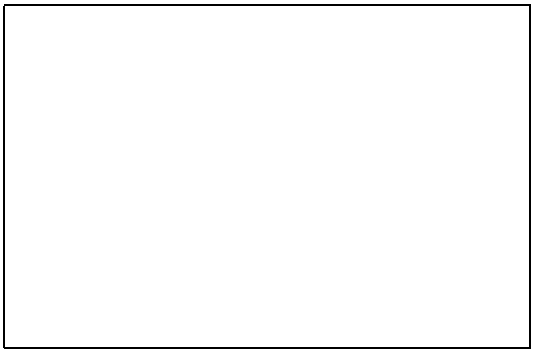


**Figure 40. Clustering with Standardized Volume & Occupancy < 26**

**Table 11 . TOD classification for V, O < 26 Cluster**



The above demonstration shows that the standardized volume and occupancy values may not be the optimal input variables to use for the cluster analysis. The volume only cluster analysis and the volume with the transformed occupancies to values < 26, result in cleaner, more refined TOD intervals. It is important to use the cleanest possible data so as not to form useless clusters. Figure 41 shows the volume and occupancy cluster centroids for the cluster analysis where the occupancies were transformed to values < 26 and then all volumes and occupancies were standardized. The error bars on this chart represent the standard deviation within the clusters. This figure shows that the centroid method using the standardized volume and occupancy < 26 values, produce clusters consistent with the TOD intervals as produced in Figure 40.

**Figure 41. Cluster Centroids and Standard Deviations**

### 5.2.2 Un-Standardized Input Variable Cluster Analyses

The cluster analysis was done on the same data set for un-standardized input variables. The resulting clusters are quite similar as those with standardized input variables. The most significant difference is that when clustering based on volumes and occupancies, the un-standardized cluster analysis does not produce the useless cluster 5 as above from the phase 6 saturated volumes. See Figure 42 for the cluster output of this cluster analysis. This is most likely due to the fact that when using raw data, the occupancies are not weighted nearly as heavily as the volumes because they lie on such a smaller scale than volume. Thus occupancies do not drive the clusters as significantly as the volumes do. Table 12 shows the TOD classifications associated with the un-standardized volume, occupancy cluster analysis. One can see that the clusters formed relate closely to those produced in the standardized volume, occupancy cluster analysis. The only differences are that here, the fifth cluster represents a PM peak plan and there is no post AM plan as

in the standardized cluster analysis. The latter may point to the fact that not as much resolution is achieved without standardizing variables; however, the occupancies will not contribute so much to producing useless clusters during times of saturation.

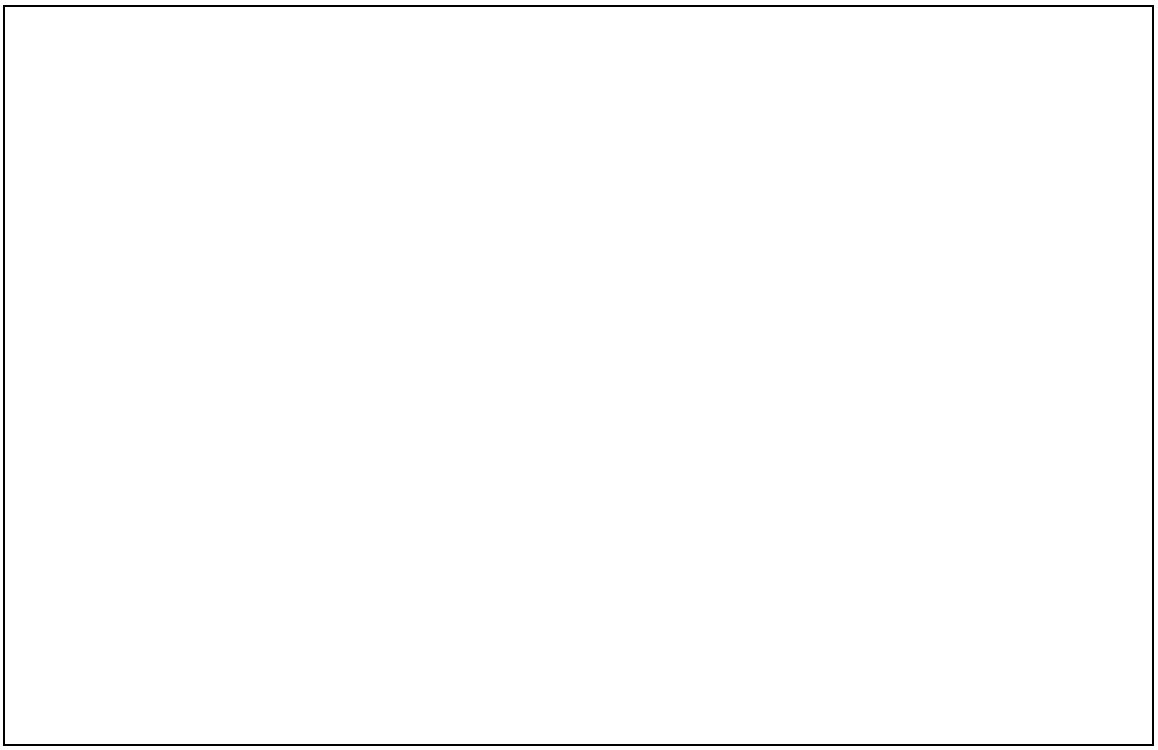
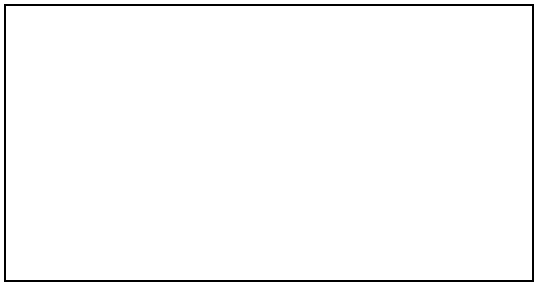


**Figure 42. Cluster with Un-Standardized Volumes and Occupancies**

**Table 12. TOD Classifications for Un-Standardized Vol & Occ Clusters**



The cluster analysis with the un-standardized volumes is nearly the same as the standardized volume cluster analysis. The standardized results for the volume only cluster analysis appear to provide cleaner TOD intervals than the un-standardized volumes.

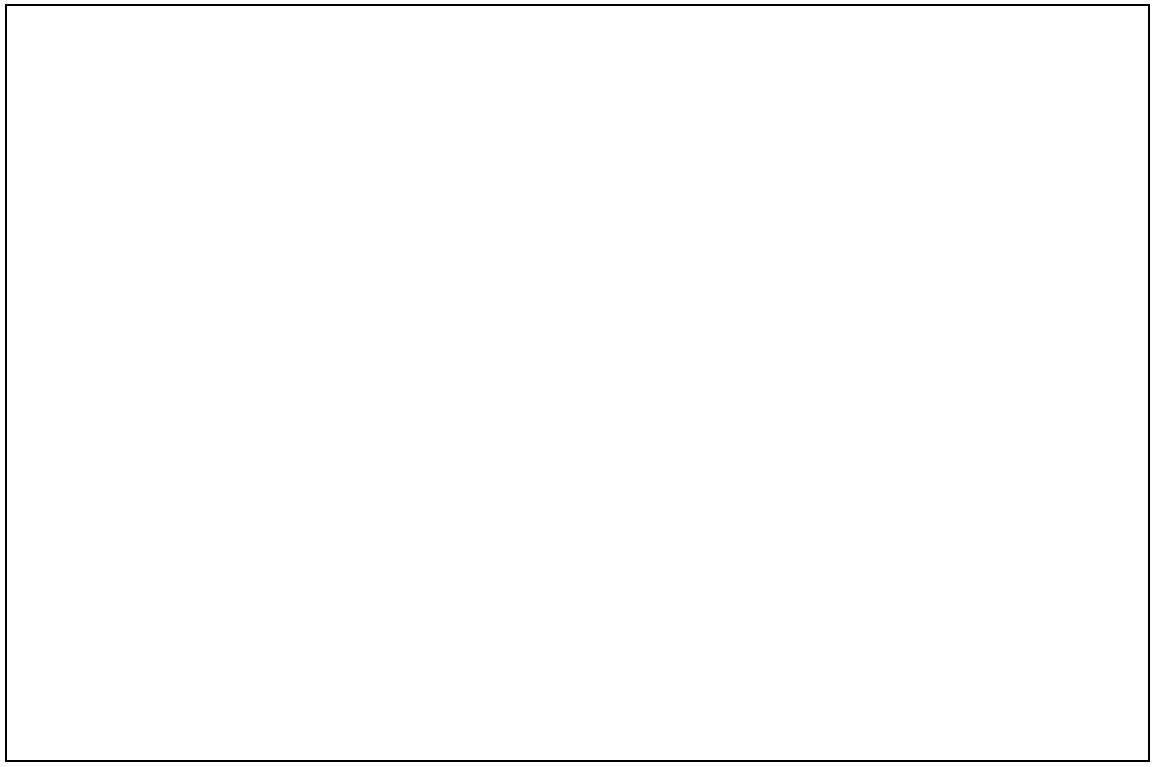


**Figure 43. Cluster with Un-Standardized Volumes**

**Table 13. TOD Classification for Un-Standardized Volume Clusters**

The cluster analyses for the standardized and un-standardized volume and occupancy < 26 analyses are also similar; however, the analysis with the standardized variables provides more refined results with more detailed TOD intervals. The un-standardized output appears to have too many transitions between clusters or timing plans without a substantial amount of observations existing for some of the transitions. For instance, there does appear to be a pre and post PM period for the un-standardized analysis, but there aren't a constant and substantial amount of observations comprising those periods, making it difficult to make such a classification. Thus, it would be recommended that standardized results provide better TOD intervals.

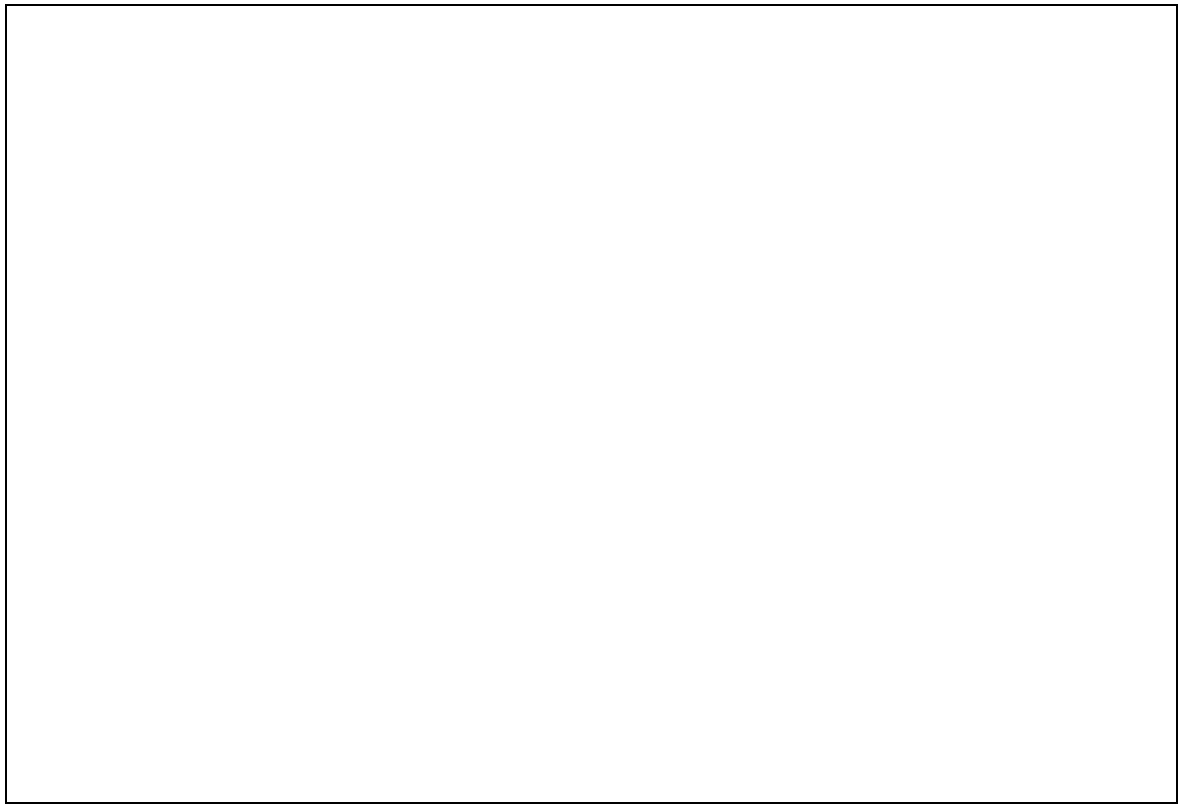


**Figure 44. Cluster with Un-Standardized Volume and Occupancy < 26**

**Table 14. TOD Classification for Un-Standardized Volume and Occupancy < 26**

For the most part, the standardized results seem to be superior except in the case where volume and occupancy are being used as input variables. To deal with such a situation, the variables could be standardized and then weighted manually so the occupancies do not contribute to the cluster formations as greatly. Otherwise, if clustering based on volumes and occupancies, it may be beneficial to use un-standardized variables to remove the emphasis from the occupancy variables. However, cluster analysis literature also recommends standardizing input variables as general practice. From the above analyses, it would be suggested that the standardized volume and occupancy < 26 variables be used in the cluster analysis in the situation that saturation occurs and large occupancies skew the results. This method responds the most effectively to the sensitivity of the changing traffic conditions throughout the day, especially during the mid-day period.

### 5.2.3 Weighted Cluster Input Variables

This section investigates the use of standardized volume and occupancy pairs, so as to keep the state definition as refined as possible, while assigning weights to the input variables to produce improved clusters. Though volume and occupancy pairs produce good results when occupancies are reduced to values of 25% or less, the weighting of volumes would eliminate the need to manipulate the occupancy data in the absense of

data cleansing tools, while retaining that information in the state definition. Figure 45 depicts the TOD intervals created by the cluster analysis with the standardized volumes and occupancy pairs where volume is weighted by a factor of 20. The 20 factor is a commonly used weighting value that represents the degree of scale difference that naturally exists between common volume and occupancy pairs (5). The TOD intervals are exactly the same as those from the standardized volume only cluster analysis and the un-standardized volume and occupancy < 26 cluster analysis. The TOD intervals appear to be cleaner in Figure 45 than in the above mentioned analyses.

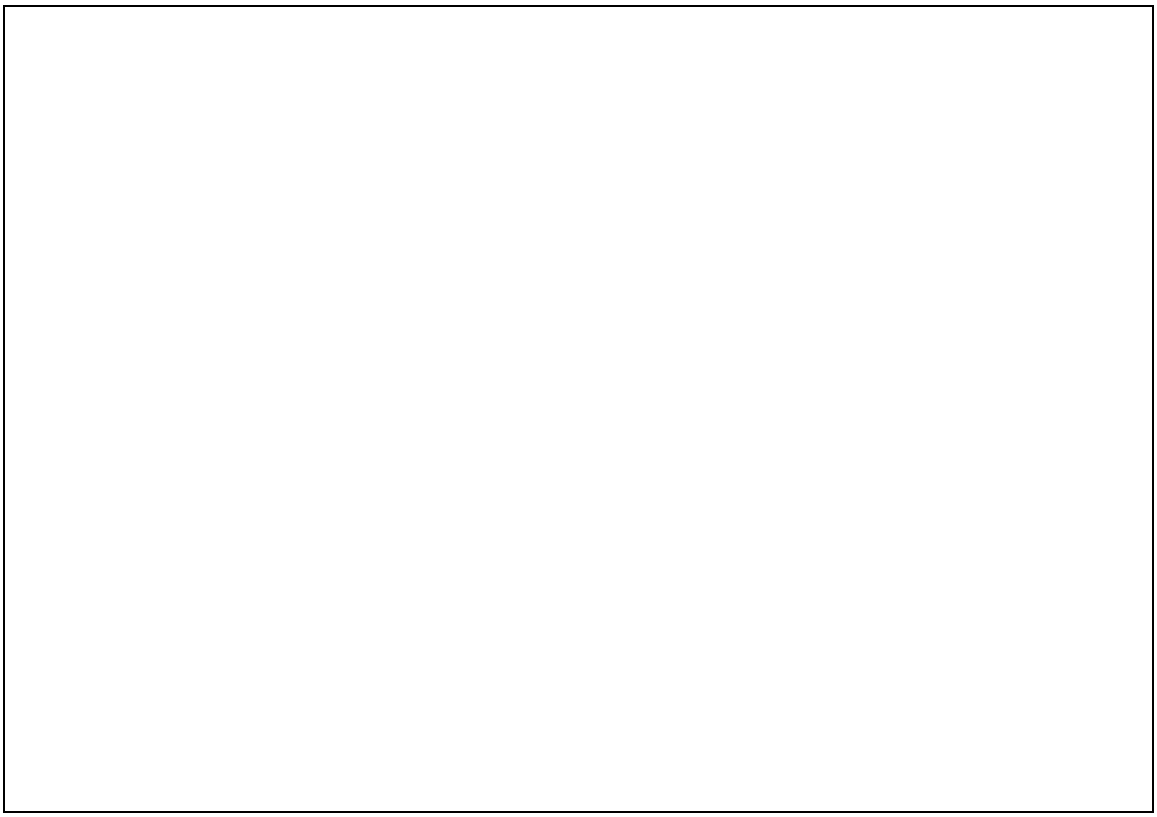


**Figure 45. TOD Intervals with Standardized and Weighted Volumes and Occupancies**

The sensitivity analysis done using alternate input variables for cluster formation shows that most combinations of cluster variables produce similar results, while some appear slightly superior to others. The worst case cluster analysis was the standardized

volume and occupancy case. This was due to the large occupancy values that don't add much information to the state definition, but rather clutter it with confusing information. When using occupancies in the state definition, values greater than 25% should be reduced to that value to reduce the possibility of meaningless cluster formations. Also, with occupancies in the state definition, the volumes can be weighted heavier. Using occupancies and volumes versus only volumes catches more of the sensitive changing traffic conditions, however both provide fairly good TOD interval results. It would be recommended from these sensitivity analyses that weighting volumes heavily or reducing occupancies to < 26 for input variables be practiced. These methods allow for good TOD intervals with standardized variables. Also, much literature exists supporting the effectiveness of volumes over occupancy in classifying traffic conditions, thus supporting conclusionary results of weighting volumes more heavily (1).

### *5.3 Sensitivity Analyses – Minimum Number of Observations Per Cluster*

One constraint imposed on the cluster analysis is that there must be a minimum number of observations existing in a cluster for it to form a unique cluster in the output. In SAS, this constraint is induced with the 'Dock = n' command, where the n is the minimum number of observations that must make up each cluster. The value of n should be dependent on the data set sample size. The use of the 'Dock' command alleviates the formation of small clusters that occur for too short a time in which it would be unsuitable to create and implement a timing plan. This sensitivity study explores the cluster output produced with different values of n, primarily with 'Dock = 1, 2' versus larger values of n. Two comparisons were done here using a 5-cluster analysis and a 6-cluster analysis. The only negative effect of imposing the 'minimum number of observations' constraint is

that it is possible to lose some observations in the cluster analysis since if they would have formed a smaller cluster than n, those observations would have been removed from the cluster output. The following cluster analyses were done with the centroid cluster methodology.

The 5-cluster analysis dock comparison looks at the use of 'Dock = 4' versus 'Dock = 2.' Figure 46 shows the TOD intervals formed when the cluster formations are constrained to containing a minimum of 4 observations. Figure 47 shows the TOD intervals formed when the cluster formations are not constrained to containing a minimum number of observations.

**Figure 46. TOD Intervals with Minimum of 4 Observations Per Cluster**

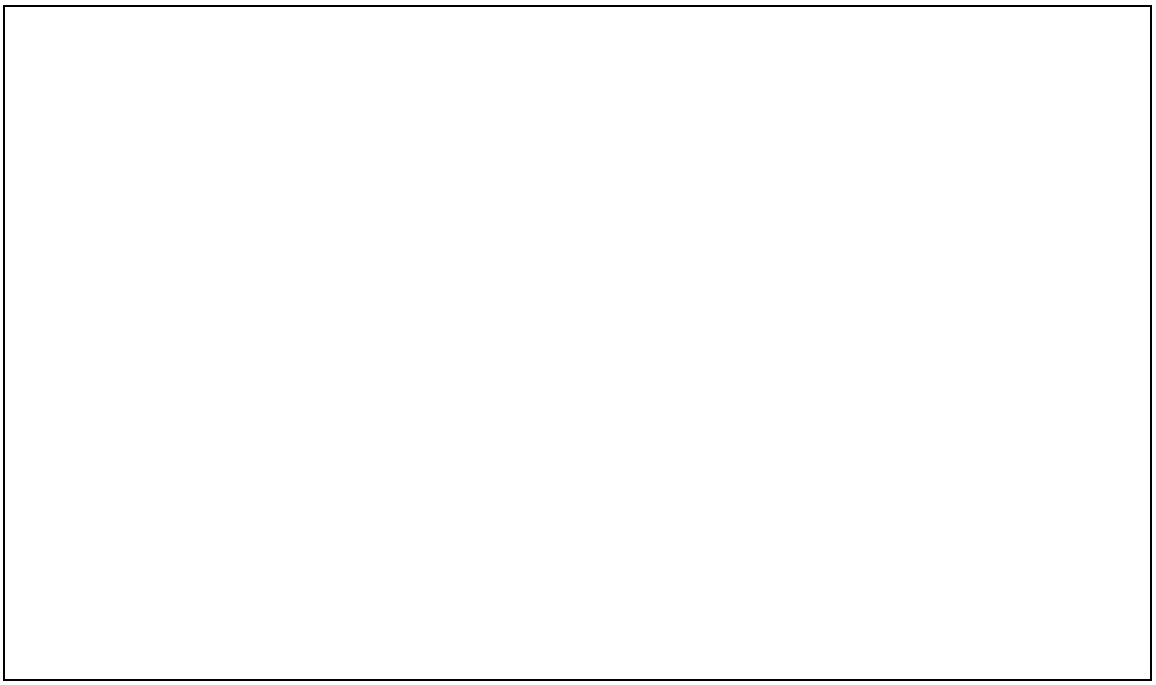


**Figure 47. TOD Intervals with a Minimum of 2 Observations Per Cluster**

The cluster analysis that does not constrain the clusters to contain a substantial amount of observations does not form very good TOD cluster intervals. The AM peak is not distinguishable, but is included in the large mid-day period. This is due to the meaningless cluster formed at the fifth level, where only two observations are contained within the cluster. Refinement of the clusters is lost when not enough clusters are formed to represent the main periods of time, as is the case in Figure 47 where the fifth cluster failed to represent a main time period due to the nearly unconstrained cluster formations. The cluster formations in Figure 46 represent clean, intuitive TOD intervals, with no wasted clusters. Analyses show that the Dock command with a value of n less than 4 produce results similar to those in Figure 47.

The next example compares the cluster outputs for 6 cluster formations with 'Dock = 1' versus 'Dock = 4.' Figure 48 represents the clusters formed from the unconstrained case. Clearly, the clusters formed here do not hold much meaning.

Clusters 4, 5 and 6 all represent a PM peak at different levels, thus leaving too few clusters to represent the remaining intervals that should exist during the day. Only two well-defined clusters exist at this point for the off peak and mid-day periods. Figure 49 shows the same cluster analysis using a constraint of at least 4 observations existing in each cluster. The clusters formed here make sense and follow an intuitive pattern. Cluster 6 is not a very solid cluster, probably due to the fact that one too many clusters were formed here. The number of clusters formed will be discussed in more detail in the next section. Also, some observations from the PM period are missing in the output due to the constraint imposed, but this is a tradeoff worth making for the formation of clean, meaningful clusters.

The studies here show that the constraint on the clustering algorithm for constraining the number of observations to a minimum value for a cluster to be formed is essential. Without this constraint, the clusters tend to form levels at which only one or two observations exist in the clusters. This commonly produced multiple levels at the PM peak period for the data set used here, as seen in Figure 48. This removes any refinement of the remaining clusters to distinguish changing traffic trends during the remaining periods of the day. One tradeoff that may be made with the use of this constraint is that some of the observations are removed from the output tree. This is due to the fact that the constraint for basing clusters on a certain number of observations removes those observation from the output that do not follow the constraint. This is apparent in Figure 49, where the PM peak observations are missing since they were excluded in the output for not containing enough observations. Yet, the clusters formed with this constraint are meaningful and catch the changing traffic conditions over the

entire 24-hour period and so this constraint should be imposed to ensure meaningful clusters. The tradeoff of missing observations with the use of the constraint does not occur with all data sets.

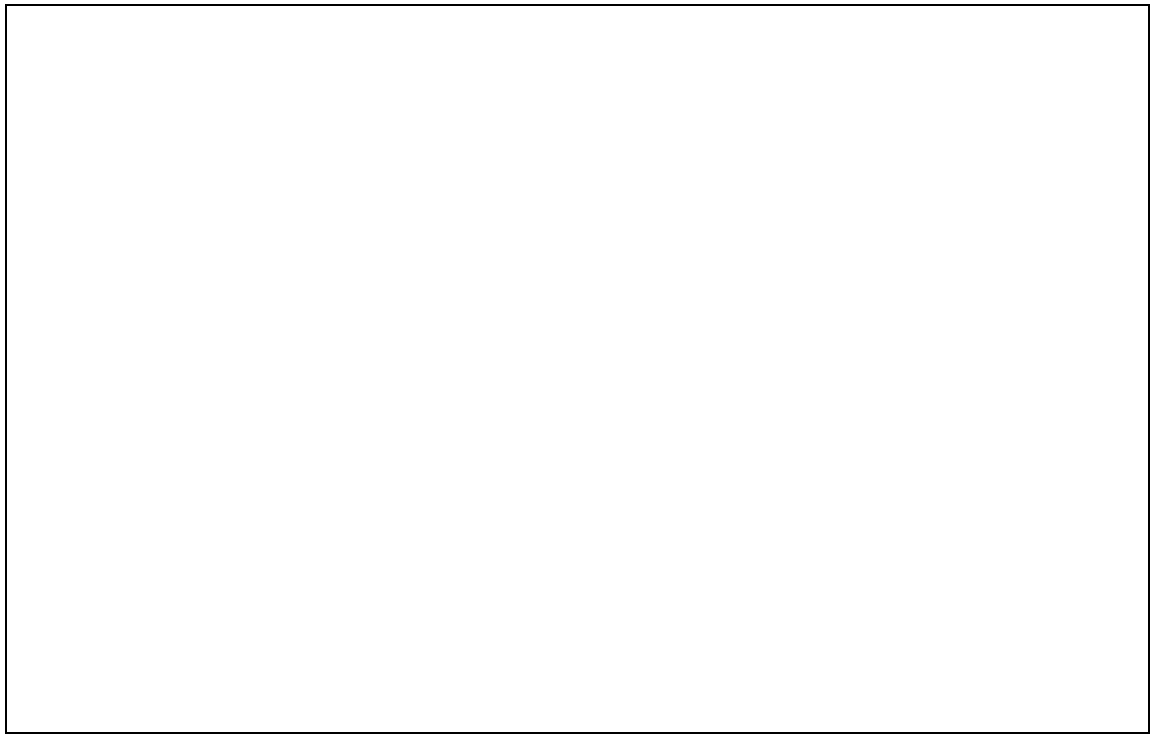


**Figure 48. TOD Intervals from Unconstrained Number of Observations Per Cluster**

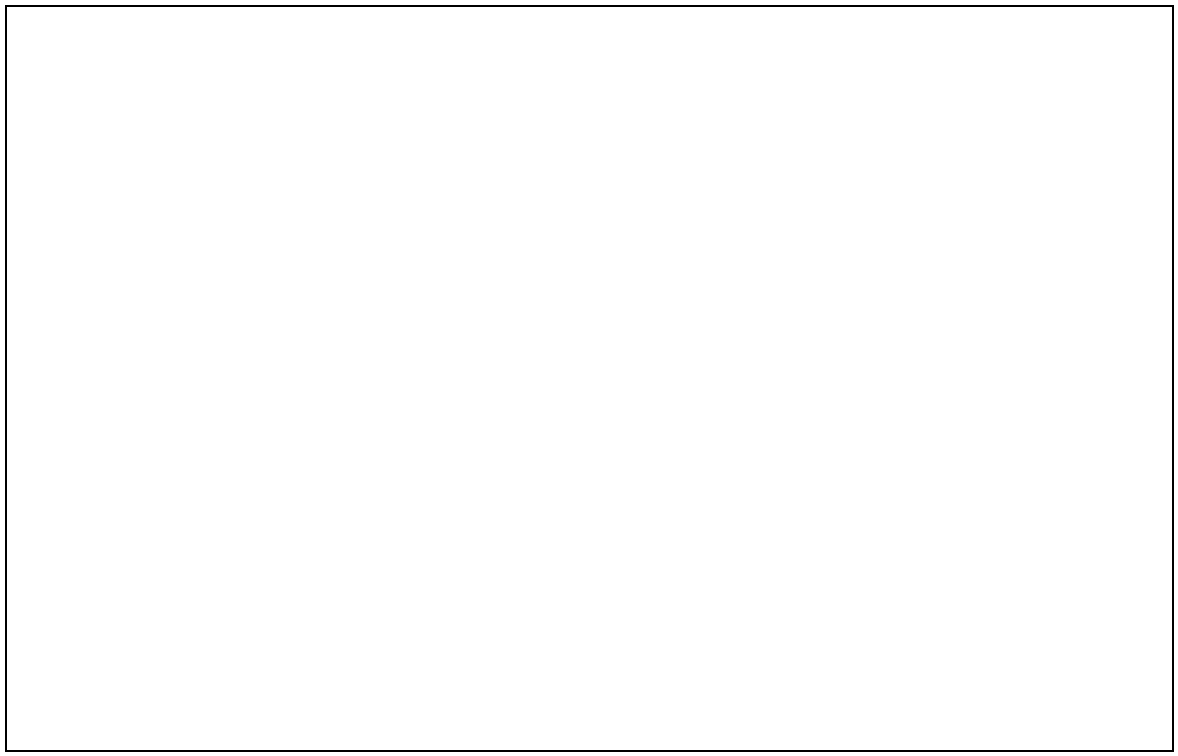


**Figure 49. TOD Intervals from Minimum of 4 Observations Per Cluster**

### *5.4 Sensitivity Analyses – Number of Clusters*

One of the most important considerations in the clustering process is the number of clusters formed. This influences the TOD intervals produced which represent the timing plans to be developed, the basis of this research. As discussed in Chapter 2, the cubic clustering criterion (CCC), the pseudo F statistic (PSF) and the pseudo $t^2$ (PST2) statistic are used for guidance in the selection of the number of clusters. This section looks at the values of these SAS statistics and the corresponding TOD interval outputs produced using different numbers of clusters. Table 15 shows the statistics for the last ten clusters formed. Of particular interest in this table are the CCC, PSF and PST2 statistics. The largest absolute value for the CCC is recommended, along with the first local maxima value of the PSF statistic and the smallest PST2 statistic. This study includes an example from the data set consisting of standardized volume and occupancy data from 8 March –

29 September 2000. Regardless of the values of the stopping statistics, the formation of 1, 2, 3 or more than 8 clusters will be ignored. The clusters are formed for representation of timing signal plans during a 24-hour period. Less than 4 clusters would not allow enough timing plans to capture the changing traffic conditions during a day and the signal controllers can only hold up to 9 timing plans at one time, at least one or two of which must contain weekend traffic plans.

In Table 15 it appears that the best choices for the number of clusters is 7 and 6 clusters is also possibly a good solution, though according to the statistics, not as optimal as 7 clusters. The largest CCC value and pseudo F statistic occur at the seventh level. The pseudo $t^2$ statistic is small, although it is a bit smaller at the eighth level, not significantly though. Cluster level 6 statistics are not as good as level 7, however they are good enough to consider the sixth level as an option. Figure 51 shows the cluster TOD intervals produced from the sixth level cluster analysis. Figure 52 shows the fifth level of cluster analysis output solely for the purpose of comparison of the outputs of an un-optimal stopping rule statistic from Table 15.

**Table 15. SAS Stopping Rule Outputs**

| NCL | Cluster History: Clusters Joined | | FREQ | SPRSQ | RSQ | ERSQ | |
|---|---|---|---|---|---|---|---|
| | | | | | | Norm Cent Dist | T i e |
| | | | CCC | PSF | PST2 | | |
| 10 | CL26 | CL22 | 6 | 0.0040 | .921 | .892 | |
| | | | 4.37 | 112 | 8.7 | 0.3752 | |
| 9 | 30DEC1899:17:00:00 | 30DEC1899:18:00:00 | 2 | 0.0017 | .919 | .884 | |
| | | | 5.11 | 124 | . | 0.3997 | |
| 8 | CL9 | CL38 | 5 | 0.0033 | .916 | .874 | |
| | | | 5.71 | 137 | 4.1 | 0.359 | |
| 7 | CL16 | CL11 | 28 | 0.0068 | .909 | .863 | |
| | | | 5.92 | 149 | 5.5 | 0.4157 | |
| 6 | CL12 | CL14 | 18 | 0.0212 | .888 | .849 | |
| | | | 3.59 | 143 | 27.1 | 0.4755 | |
| 5 | CL6 | CL7 | 46 | 0.0475 | .841 | .830 | |
| | | | 0.79 | 120 | 28.8 | 0.4536 | |
| 4 | CL13 | CL10 | 36 | 0.0269 | .814 | .803 | |
| | | | 0.74 | 134 | 33.5 | 0.5051 | |
| 3 | CL5 | CL15 | 55 | 0.0507 | .763 | .757 | |
| | | | 0.24 | 150 | 21.3 | 0.5659 | |

Figure 50 shows the TOD intervals produced by the cluster analysis at the seventh cluster level, which should represent the optimal clustering according to statistics. The TOD intervals formed here are clean clusters that occur at intuitive times of day. There exists a clear off peak, an AM peak, a post-AM peak, a mid-day peak, a PM peak, a post-PM peak that returns to the mid-day peak cluster, and then two transitions during the evening period before returning to off peak. The intervals are refined, and the consideration of too many transitions should be considered here.

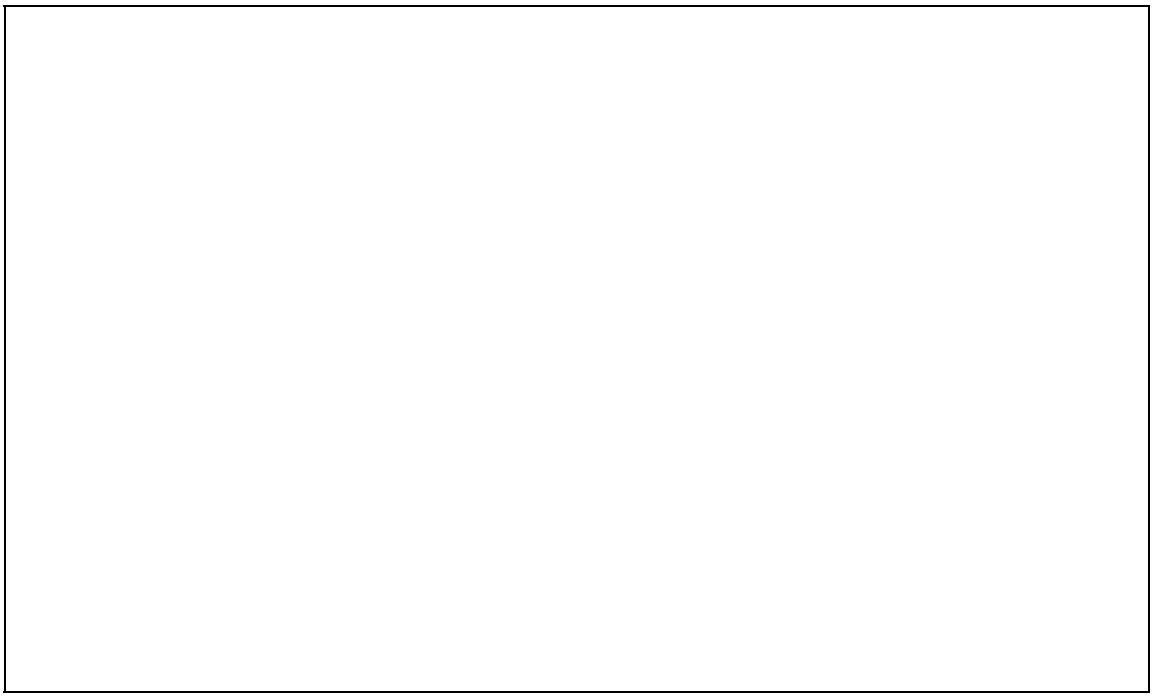


**Figure 50. Optimal Number of Clusters (7 Clusters)**

Figure 51 shows the TOD intervals for the sixth cluster level. The only difference between these TOD intervals and those formed from the seventh cluster level is that the AM and post-AM peak have merged into one cluster. This signifies that those two timing periods probably consisted of the most similar traffic conditions. This result shows a

little less refinement in the production of TOD intervals, however this may not be a significant effect for the production of timing plans if those two plans were similar enough.

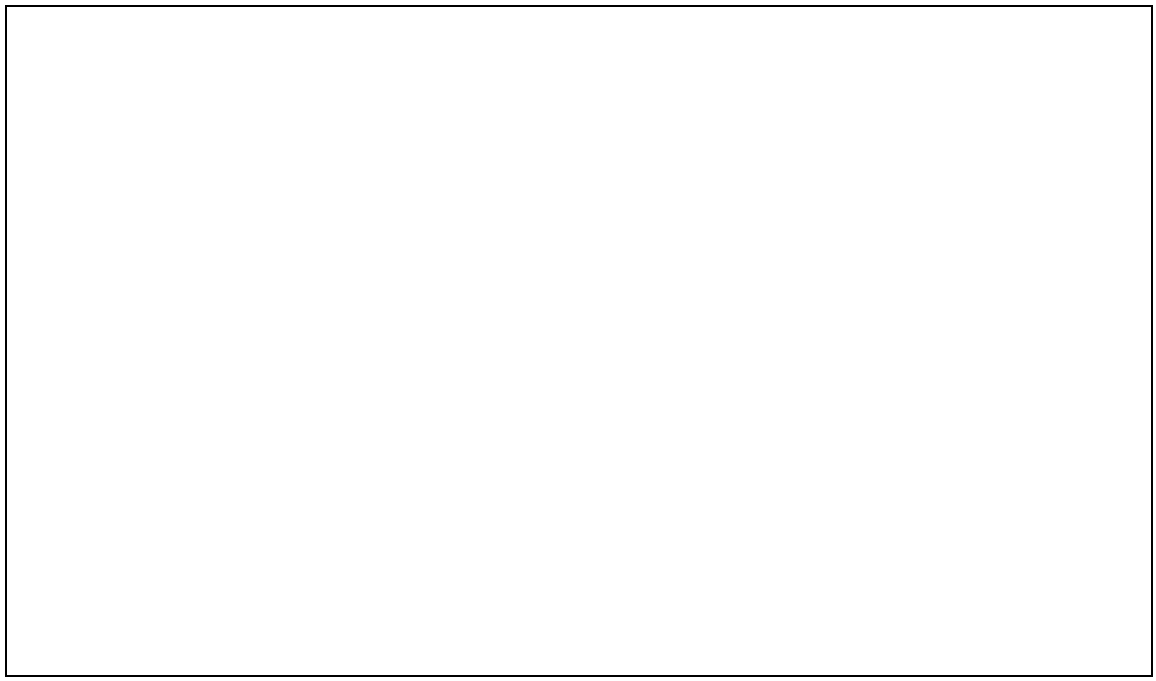


**Figure 51. Optimal Number of Clusters (6 Clusters)**

Figure 52 shows the cluster outputs from the fifth cluster level. From Table 15, these stopping rule statistics are not as optimal as those are for the sixth and seventh cluster levels. This output is the same as that for the sixth cluster level, except that now the AM peak is merged into the mid day peak period. This would probably have a more significant effect on signal plan development, since the AM peak period is typically an important plan for a 24-hour period. This shows that the CCC, PSF and PST2 values are accurate descriptors for choosing the number of clusters and the choice of number of clusters should produce the most refined and intuitive results when following these stopping rules.

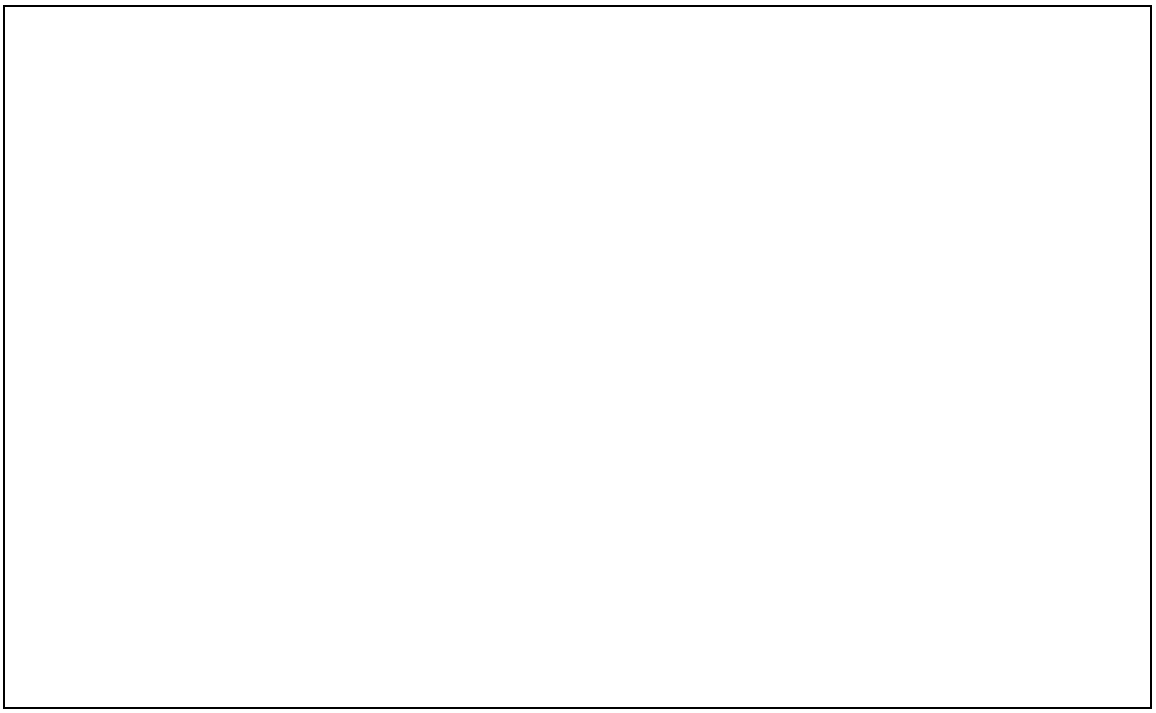


**Figure 52. Optimal Number of Clusters (5 clusters)**

### *5.5 Single Intersection – Baron Cameron & Reston Parkway Case Study*

The single intersection case study is from the Reston Parkway and Baron Cameron intersection. See Figure 11 for the layout of this intersection. The single intersection case study was performed prior to the three-intersection corridor case study to ensure a complete process with valid results on a simple case. The results are presented here to make a recommendation for this process for single intersections as well as corridors and to support the claims made for the proposed procedure improving corridor performance. The analysis of the results for the single intersection is brief, due to its minor role in this project. Figure 53 shows the TOD intervals developed from a cluster analysis with 4 clusters being the optimal number of clusters for this data set. Table 16 displays the times associated with the new TOD intervals taken from Figure 53. This table also shows

the old TOD intervals, which are similar, but differ the most in the shortened periods for the new intervals. Unique timing plans were developed for the off peak, AM, MD and PM periods. This is the same four timing plans developed implemented currently by VDOT, however the cluster analysis does transition between them more than once. This is the case for cluster 2, which represents a pre-AM, MD and post-PM period. This clustering across times-of-days is intuitive in that the traffic conditions represented at these opposing times would probably assume similar traffic states.

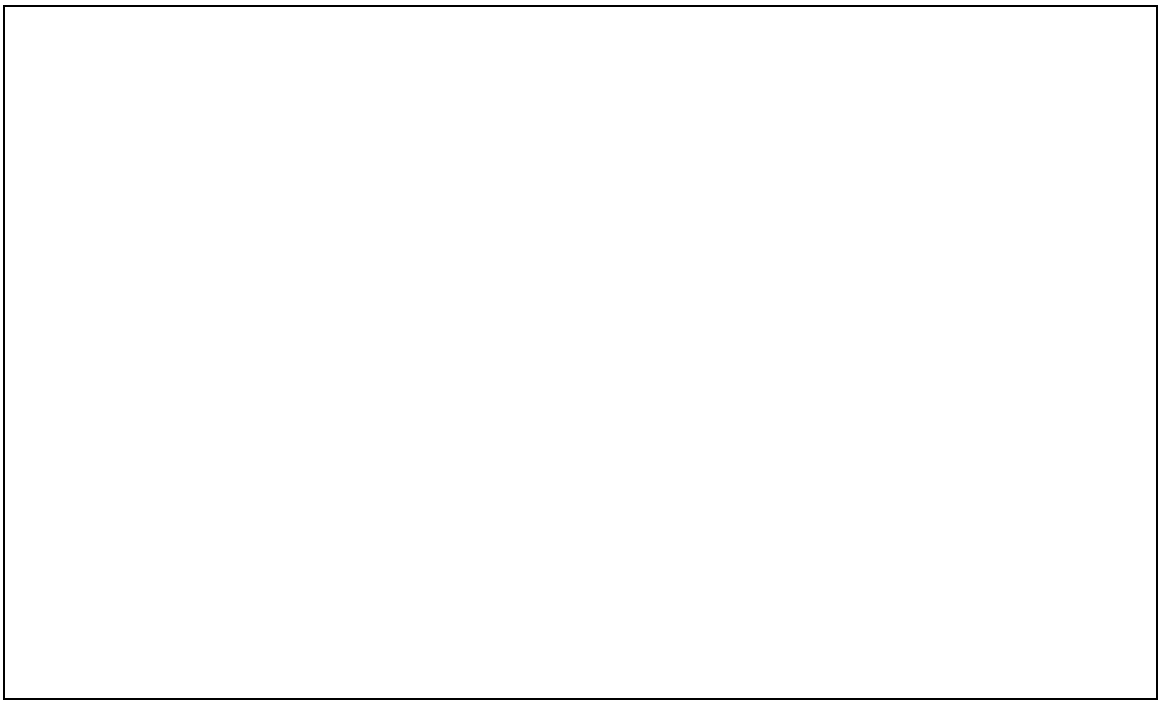
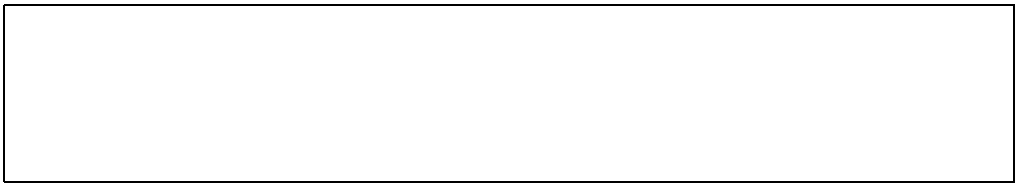


**Figure 53. TOD Intervals at Baron Cameron & Reston**

**Table 16. TOD Interval Classification for Baron Cameron**



The same four scenarios are examined for the single intersection case as for the 3-intersection corridor case. These scenarios include the performance of:

1. The old plans & old TOD's
2. The old plans & new TOD's
3. The new plans & new TOD's
4. The new plans & old TOD's

Figure 54 shows the simulation MOP results from the single intersection case. These results vary from those of a multiple-intersection coordinated system in that the old TOD's perform significantly better than the new TOD's. This may possibly be due to the fact that since the TOD intervals are similar to the old ones, the increased transition effects dominate the increase in transitions. This also demonstrates the ease in identifying TOD intervals for single intersections, since the intervals can be based solely on that single intersection without any concern for corresponding intersections in the system. The difficulty in TOD interval selection arises as more intersections become involved in a coordinated system, since manual identification does not take into account traffic conditions at every intersection and every movement. Thus the traffic engineers rely on the critical intersection demand for TOD intervals and the remaining intersections are not considered. It can be hypothesized that as the corridors become more complicated with more intersections, the automation of TOD interval selection would be increasingly significant in identification of optimal intervals. The newly developed timing plans, however, perform significantly better than the old timing plans as is the case for the corridor case.

**Figure 54. Simulation Outputs from Single Intersection**

The single intersection case study supports the use of timing plan development with system detector data versus hand-counted data. It also shows simplicity of selecting TOD intervals manually at single intersections and supplies a basis to the theory that the increasing difficulty of TOD selection is due to the increase in intersections that do not contribute to the selection of the TOD intervals in corridors.

### *5.6 Three-Intersection Corridor Case Study Results*

The three-intersection corridor case study includes the intersection of Reston Parkway with Sunset Hills, Bluemont and New Dominion. See Figure 12 for the corridor layout of these intersections. The clusters for this case study were validated in Chapter 4 and the simulation results can be viewed as an external validation of the cluster formations based on the performance on traffic conditions of the resulting clustered timing plans. Figure 55 shows the TOD intervals developed from a cluster analysis with 7 clusters. Table 17 displays the times associated with the intervals taken from Figure 55. Unique timing

plans were developed for the off peak, AM, post-AM, PM, evening and pre-Off periods. The seventh timing plan represents the mid-day as well as the post-PM periods.

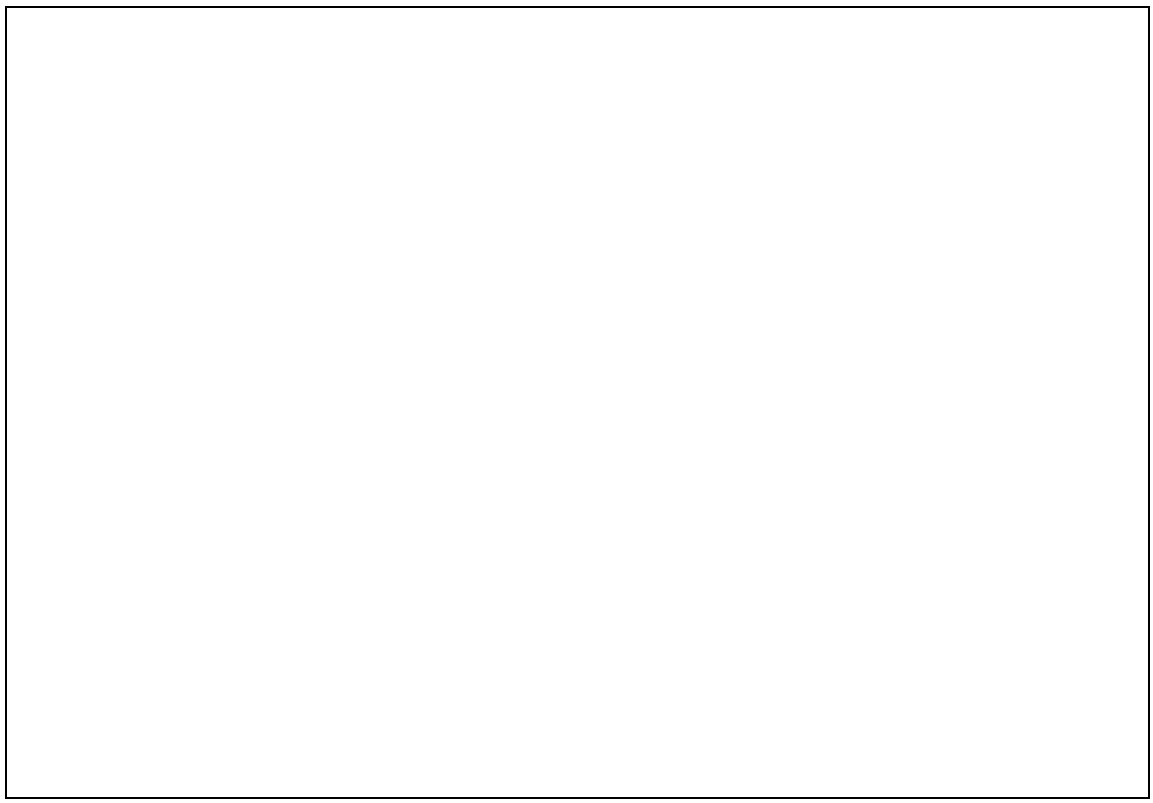
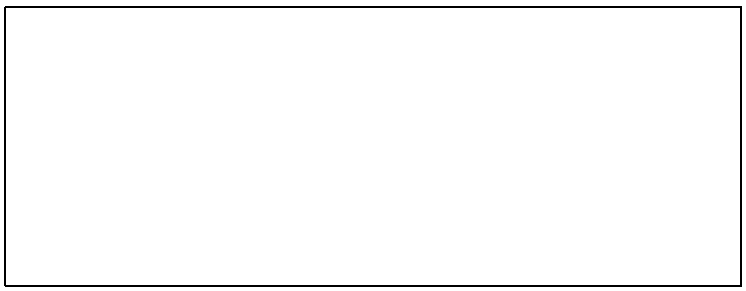


**Figure 55. TOD Intervals for 3-Intersection Corridor**

**Table 17. TOD Interval Classifications for 3-Intersection Corridor**



This case study supports the hypothesis that timing plans can be improved through the use of data mining tools. Figure 56 shows the outputs of the simulation from four scenarios. These scenarios include the performance of:

1. **Old Plan, Old TOD** – Plan developed with hand-counted volumes, implemented during the handpicked TOD intervals based on critical intersection traffic.
2. **Old Plan, New TOD** – Plan developed with hand-counted volumes, implemented during newly clustered TOD intervals based on full state definition.
3. **New Plan, New TOD** – Plan developed with database volumes from 6 months, implemented during newly clustered TOD intervals based on full state definition.
4. **New Plan, Old TOD** – Plan developed with database volumes from 6 months, implemented during newly clustered the handpicked TOD intervals based on critical intersection traffic.

Detailed comparisons of these four scenarios follow. "Old Plans" refers to the timing plans developed by VDOT from the one-day, hand-counted volumes. "New Plans" refers to the timing plans developed in Synchro from the historical database volumes. "Old TOD's" refers to the time-of-day intervals selected by VDOT engineers based on the critical intersection and intuition. "New TOD's" refers to the time-of-day intervals produced from the cluster analysis, where the newly developed intervals are based on all intersection and movements in the corridor for more refined intervals. These four headings will be found in charts and analyses to follow and comprise the four scenarios being studied.

The four main simulation outputs used to evaluate performance of opposing timing plans are:

- Travel time
- Total delay
- Fuel used
- Denied entry

These measures of performance are defined in detail in *Section 3.11.1.1*.

From Figure 56, it can be seen that the new plans & new TOD's do in fact perform better, while the current plans implemented, represented by the label 'Old Plans,

Old TOD's,' perform the worst. The combination of old plans with new TOD's and new plans with old TOD's fall in between the two extremes as will be discussed in the following section. The current plans, which form the basis of comparisons in the analysis, are the plans developed by VDOT. These plans were recently optimized approximately one year ago with hand-counted volumes. These plans are considered "newly updated," by the VDOT traffic control center in northern Virginia. Since the northern Virginia arterial network consists of approximately 120 corridors, the Reston Parkway corridor will not again be updated for many years. Through interviews conducted with the traffic control engineers, it was learned that the traffic engineers spend periods of weeks, or even months in re-optimizing timing plans for one corridor. By the time this process is completed for one cycle of the all the corridors, quite a few years may go by before the cycle can be restarted for another re-optimization of plans at each corridor.

**Figure 56. SimTraffic Outputs for Three Intersections**

The four scenarios that were simulated have been shown to display significant amounts of variance between the scenarios. This demonstrates the significant variance between the measures of performance for the different scenarios (old plan & old TOD, new plan & new TOD, old plan & new TOD, new plan & old TOD). F-tests were conducted for each of the four measures of performance: travel time, delay, fuel used and denied entry, to measure the between-scenario variation to the variation calculated from within each scenario. The F-statistic deals with ***I*** populations (scenarios) with a random sample of ***J*** observations from each one (28). The F-test is valid under the assumptions that the distribution of the ***I*** scenarios are normal with the same variance. The F distribution arises from a ratio in which there is one number of degrees of freedom (df) associated with the numerator and a different df associated with the denominator. The variable $v_1$ and $v_2$ denote the degrees of freedom associated with the numerator and denominator respectively. For the MOP variance testing, the parameters are as follows (28):

**F = MSTr / MSE**, where

MSTr = between-sample variation

$$MSTr = J / (I - 1) \Sigma_i (\mathbf{X}_i - ((\Sigma_i^I \Sigma_j^J \mathbf{X}_{ij}) / IJ))^2$$

MSE = within-sample variation

$$MSE = S_1^2 + S_2^2 + \ldots. + S_I^2 / I$$

$$Si2 = \Sigma_{j=1}^J (\mathbf{X}_{ij} - \mathbf{X}_i)^2 / J - 1$$

$$V1 = I - 1$$

$$V2 = I ( J - 1)$$

I = 4 scenarios

J = 6 observations in each scenario

The null hypothesis ($H_0$) being tested is that the means between scenarios are equal. The null hypothesis can be rejected if (28):

$\mathbf{f \geq F_{\alpha, I-1, I(J-1)}}$, where

$\alpha = .05$
$H_0 : \mu_1 = \mu_2 = \ldots = \mu_I$
$H_a$: at least two of the $\mu_i$'s are different

Table 18 shows the computed f, compared to $F_{\alpha, I-1, I(J-1)}$, and the p-value, which represents the level of significance with which the null hypothesis can be rejected. As long as the p-value remains less than the testing level of significance, .05, then it is assumed that the null hypothesis should be rejected.

**Table 18. F-tests across 4 scenarios**

| | |
|---|---|

From this table, it is clear that the null hypothesis for each MOP is rejected. This implies that at least two of the scenario means are significantly different. To determine which of the scenarios are significant from each other, paired t-testing will follow in *Section 5.5.4*.

### 5.6.1 Three-Intersection Case Study Assumptions

The assumptions made for the process of developing and simulating new timing plans are stated in the following list.

- The system detector data accurately represents actual traffic conditions occurring in northern Virginia, thus the simulations are based on an accurate representation allowing for validation of the newly developed timing plans.
- The volume and occupancy data are normally distributed over similar times-of-day, thus validating the averaging technique used to create one representative value for each TOD. See Figure 20 for a normally distributed time-slice of data.

- The data representing each of the simulation MOP's for each simulation run are normally distributed, thus validating the use of t-tests for significance testing.

SimTraffic accurately models transition effects as experienced in the field and does not conceal declined performance due to increase transitions.

### 5.6.2 Evaluation of Simulations

Refer to Section *3.11.1.1* for full detail of the measures of performance described briefly here. The Travel Time is a total of the time each vehicle was present in the simulation area. The travel time includes time spent by vehicles Denied Entry. Total Delay is equal to the travel time minus the time it would take the vehicle with no other vehicles or traffic control devices (16). Total delay also includes all time spent by denied entry vehicles while they are waiting to enter the network. Fuel Used is calculated with the fuel consumption tables. The fuel used in each time slice is determined by the vehicle's fleet (car, truck, or bus), speed, and acceleration. Denied entry is one of the most important measures of performance because it is a measure of vehicle throughput in the system. Denied Entry is a count of vehicles that are unable to enter a link due to congestion. Denied Entry can also be used to determine the Network Throughput. In a congested network, lower values of Denied Entry indicate increased throughput and vice versa. This is a good determining factor for the effectiveness of timing plans. The higher the number of denied vehicles typically infers that those timing plans are performing worse. The calculations of these MOP's are fully detailed in Chapter 4. The totals of these MOP's are a function of the number of cars in the system for each scenario. The number

of simulations run was evaluated to be significant at 6 runs for equalization of the number of vehicles across scenarios; this is explained below in *Section 5.5.2*.

### 5.6.3 Number of Simulation Runs

The current practice by VDOT for testing timing plans with simulation is to run three simulations with different random number seeds to ensure stable results. This research investigated the stability of running more than three simulations to ensure accurate simulation results. The number of vehicles that enter the system is the stabilizing variable that should equalize for each simulation scenario. T-tests were performed on the 'Vehicles Entered' variable to ensure the means were equal between different scenarios at the $95^{th}$ confidence level. For all of the hypothesis testing done in the analysis, it can be assumed that the data is normally distributed about the mean. Table 19 shows the t-test results for the comparison of the four scenarios used to evaluate timing plan effectiveness. The null hypothesis tested is that the means of the two samples are equal and since the null hypothesis is not rejected for any of the scenarios, the means are assumed to be equal, thus validating that 6 simulation runs is sufficient for producing accurate results.

**Table 19. t-test Results for Number of Simulation Runs**

### 5.6.4 Improvements with New Plans

The use of data mining tools to aid in timing plan development can benefit two aspects of the process. The first aspect is the development of new timing plans based on 90$^{th}$ percentile volume data retrieved from the historical database. The procedure would replace the current practice of hand-counting cars to develop timing plans. The second aspect is to look at the refined TOD intervals, which are developed with cluster analysis based on similar traffic conditions occurring over the course of a day. This method would allow for a data driven selection of TOD intervals rather than an intuitive, human selection based on the aggregate volumes at the critical intersection. The analysis will be broken down into these two parts to provide a sense of where the most gains are achieved; through the new plans or through the new TOD intervals.

To visualize the benefits of the newly developed plans based on 90$^{th}$ percentile volumes from the database, a chart of the percent reductions from the new plans over the old plans is displayed Figure 57. The old plans that these reductions are being compared to are the currently implemented plans where the volumes were achieved through the one-day, hand-counted process. The first bar in Figure 57 represents the gain of the new plan over the old plan for the old TOD intervals, while the second bar represents the gain of the new plan over the old plan for the new TOD intervals. Both of these comparisons are being made to show the reductions from the use of the new plans over the old plans, while holding the TOD intervals constant. The third, dotted bar represents the gain that would be achieved from the use of the new plan and new TOD intervals over the current plans being implemented (the old plans and the old TOD's).

**Figure 57. MOP Gains of New Plan over Old Plans for Old and New TOD's**

All of the gains achieved for delay, travel time, fuel used and denied entry of new plans over old plans, evaluated during the old TOD intervals are significant at the 95$^{th}$ confidence level. The null hypothesis tested was that the means of the old and new plans were equal. The t-tests are based on 6 simulation runs and the statistics are displayed in Table 20. The t-test results for the new plans vs. the old plans, evaluated for the new TOD intervals, as outlined in Table 17, are displayed in Table 21. These results show that at the 95$^{th}$ confidence level; delay, travel time and denied entry are significant improvements for the new plans over the old plans; however, fuel used is not a significant gain, with the new TOD intervals. However, with the old TOD intervals, all MOP's improve significantly when the new plans are compared to the old plans.

**Table 20. t-test Results for New Plans vs. Old Plans Evaluated at Old TOD Intervals**

**Table 21. t-test Results for New Plans vs. Old Plans Evaluated at New TOD Intervals**

### 5.6.5 Improvements with New Time-of-Day Intervals

To visualize the effects of the newly developed TOD intervals from the cluster analysis over the old TOD intervals, the percent reductions can be viewed in Figure 58. The first bar represents the gain of implementing the old plans during the new TOD's over the old TOD's. This bar is presented in Figure 58 for a full comparison of the scenarios; however, it is unrealistic since the old plans would never be implemented over new TOD intervals. If the new TOD intervals were developed, then the new volume plans would automatically be produced in correspondence with the new TOD intervals and these would be used together (As seen in bar three in Figure 59). The second bar represents the gains achieved from implementing the new plans for the new TOD's versus the old TOD's. Here the timing plan is held constant for each of the two comparisons while the opposing TOD intervals are compared. Less lift is achieved when only adjusting the TOD intervals versus renewing the plan itself as displayed in Figure 57. The third bar

again represents the percent gain for each of the MOP's of implementing both the newly developed timing plans at the new, clustered TOD intervals over the current plans (old plans and old TOD's). This is the case where the optimal amount of lift is achieved and is displayed to show that a smaller proportion of gains come from using the new TOD intervals over the old TOD intervals, while the larger proportion of lift is coming from the new plans.

**Figure 58. Percent Gains of New TOD's over Old TOD's for Old Plans & New Plans**

The improvements of implementing timing plans over the new TOD intervals versus the old TOD intervals is not as significant as when the newly developed plans are also implemented. Table 22 shows the t-test results of the significance of the improvements for implementing the old plan during the new TOD versus the old TOD. These results show that at the $95^{th}$ confidence level, none of the MOP's measured provide significant amounts of improvement. This scenario is not as accurate as the TOD comparison for the new newly developed timing plans since for the old plans, the exact plans had not been developed for the new TOD intervals. Also, this scenario is

unrealistic, as it would never be implemented in such a combination in the field. For the second bar, the case where new plans were implemented over the new TOD intervals versus the old TOD intervals, only delay resulted in a significant improvement. Fuel used actually provided negative gain to the new TOD intervals. These results can be viewed in Table 23. It is possible that the increase in fuel usage for the new plans and new TOD's is due to the fact that with increased throughput in the corridor, vehicles are able to travel at higher speeds and accelerate faster. Since the fuel used is calculated based on speed and acceleration, this may cause an increase in fuel usage. Also, the types of vehicles in the system effect the fuel usage calculation. It is possible that more trucks and busses were present in the system, causing a higher fuel usage value for the new plans and new TOD simulations.

Table 24 shows the t-test results for the third bar that is displayed in both 'Percent Gains' charts. These are the improvements achieved using both the new timing plans and the new TOD intervals versus the old timing plans and the old TOD intervals, which are those currently implemented in the field. This third bar is the same as that in Figure 63. All of these MOP's provide significant improvements at the 95$^{th}$ confidence level.

**Table 22. t-test Results for New TOD vs. Old TOD Intervals Evaluated by Old Plans**

**Table 23. t-test Results for New TOD vs. Old TOD Intervals Evaluated by New Plans**

**Table 24. t-test Results for Old TOD & Old Plan vs. New TOD & New Plan**

### 5.6.6 Time Periods where New TOD Intervals show Significant Improvements

Due to the fact that only delay provided significant improvement when implementing the new TOD intervals over the old TOD intervals, a 24-hour breakdown shows the periods during which significant improvements are achieved. Narrowing the confidence intervals would have supported a significant improvement in all the MOP's, however a $95^{th}$ confidence level is preferable, so the periods of the day at which this is achieved are displayed in Figure 59. From 17:45 – 24:00, the new TOD intervals for the new plans performed significantly better than the old TOD intervals for the new plans. From the figure below, it is clear that a portion of the current PM peak period is significantly improved with the clustered TOD intervals. The remainder of the day (00:00 – 17:45), performed at similar rates of performance between the new TOD intervals and the old TOD intervals, both with the new plans. Gains during the peak periods are critical and the most effective since those are the times where the most difficulty is met by traffic engineers in increasing flow through the corridor. From 17:45 – 24:00, is also the portion

of the day in which the majority of the new plan periods were developed by the cluster analysis excluding the post-AM period. Figure 55 shows the clustered TOD intervals produced for the 3-intersection corridor.

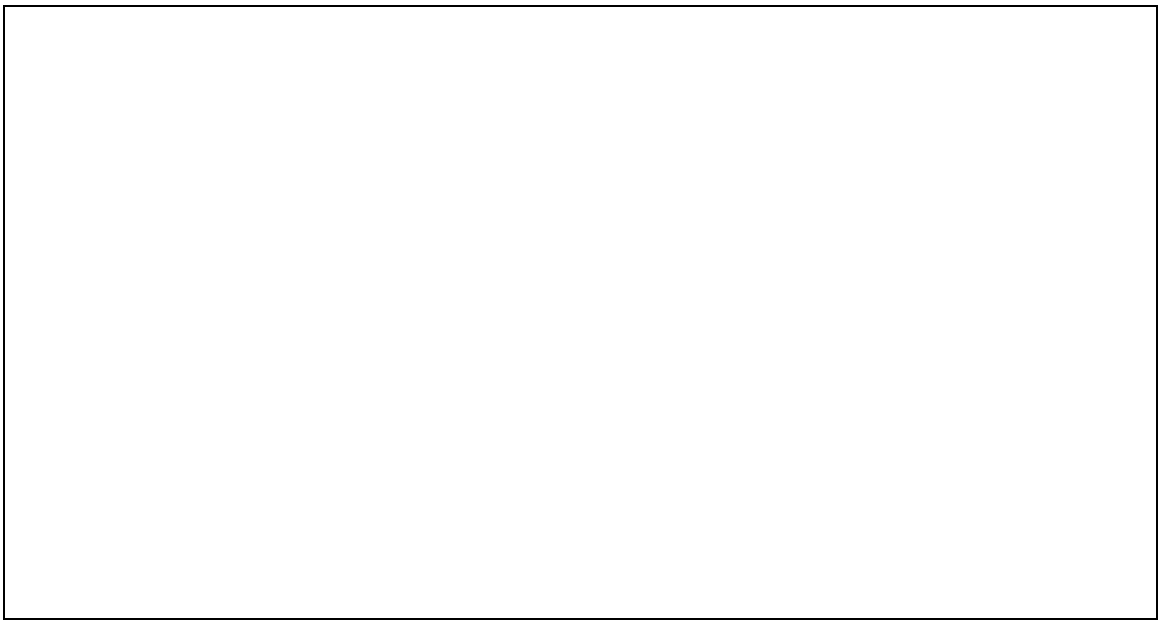


**Figure 59. Periods of Significant Gains from New TOD's versus Old TOD's**

The delay/vehicle over the 24-hour period is discussed on page 154, where the delay can be seen to be much lower during this time period for the new TOD's and new Plans versus the Old TOD's and New Plans. The fact that the significant improvements do in fact occur during the newly developed plan periods, supports the hypothesis that clustering produced refined TOD intervals, better suited for traffic conditions through a corridor.

Again, the level of significance for the t-tests on the new TOD's versus the old TOD's is at the .05 level. The null hypothesis is that the means of these two samples are

equal. Table 25, Table 26 and Table 27 show the results of the t-tests for comparing new TOD's versus old TOD's during the 17:45 – 24:00 period.

**Table 25. PM - Post PM, t-test Results for New vs. Old TOD Intervals**

**Table 26. Post PM - Evening, t-test Results for New vs. Old TOD Intervals**

**Table 27. Evening - Pre-Off - Off, t-test Results for New vs. Old TOD Intervals**

### 5.6.7 Volumes from Old Timing Plans vs. New Timing Plans

Under the assumption that the historical data base values represent actual traffic conditions on the roadway, the following charts represent the volumes used in opposing timing plans with the actual traffic conditions. Figure 60, Figure 61 and Figure 62 show these plots for each intersection in the three-intersection corridor. The 'UTDF

VPHPmvmt' data label represents the averaged historical volumes extracted from the database to represent actual traffic conditions. These are the volumes used in the simulation. The '90%, New TOD' data label represents the 90$^{th}$ percentile volumes taken from the clustered volumes for newly developed timing plans corresponding to the new TOD intervals. The movement of the volumes across the 24-hour period in the following figures shows the transitional points from plan to plan. Periods of equal volume represent opposing times of the day that operate under the same plan. The '90%, Old TOD' data label represents the 90$^{th}$ percentile volumes taken from the historical database for plan development based on the old TOD intervals, which are the same as those being currently implemented by VDOT. The 'Original Volumes' label represents the volumes used for timing plan development by VDOT, which are the current timing plans implemented at these intersections. These TOD intervals match those from the '90%, Old TOD' data since both of these lines represent transitioning through the timing plans based on the current interval selection method. It is clear that the original volumes, which are the hand-counted volumes for timing plan development, are much too low to handle the traffic conditions that actually exist. Timing plans should be developed for the high end of traffic volumes to ensure enough green time during the most congested periods of TOD intervals. Hence the use of the 90$^{th}$ percentile volumes for timing plan development. To account for the heaviest periods under the current means of hand-counting cars, the traffic engineers count during what they assume to be the peak 15-minute period of a timing period. The two newly developed plans account for the high end of traffic volumes during the timing period that plan is implemented. These plans are naturally better suited to handle actual traffic conditions and the TOD intervals are

tailored to the changing traffic trends. A one-day hand count for timing plan development is not a reliable measure as is clear from the charts below.

**Figure 60. Timing Plan Volumes versus Actual Volumes at Sunset Hills**

**Figure 61. Timing Plan Volumes versus Actual volumes at Bluemont**

**Figure 62. Timing Plan Volumes versus Actual Volumes at Bluemont**

Another conclusion that can be drawn from these volume figures deals with the selection of TOD intervals. At the three intersections representing the corridor, the above figures show that different volumes occur at opposing intersections at similar times of day. For this case study, the largest volumes occur at Sunset Hills (the critical intersection in the Reston Corridor) in the 9000 VPHPmvt range, while the smallest volumes occur at Bluemont in the 5000 VPHPmvt range. These opposing volumes occur during the same peak periods, with the maximum volumes shifting only slightly in time. Of course the peak volume shift would become more severe as the size of the corridor were increased. The current method of TOD interval development is based on the critical intersection alone, since that is the intersection servicing the maximum amount of traffic. The number of vehicles, occurring at cascading times through the corridor, including all turning movements and directions cannot be taken into account with manual TOD interval selection. These charts show that traffic does follow different patterns at

coordinated intersections, even those right next to each other and it may be beneficial to consider all intersections versus one in the selection of TOD intervals.

### 5.6.8 Gains of New Plan versus Current Plans

Any of the combinations of the newly developed timing plans and/or TOD intervals will increase performance in a signalized corridor. This section looks at the comparison of these combinations of new timing plans to the original plans, which are those being used currently with the old plans and old TOD's. Figure 63 shows the percent gains achieved from the three combinations of new scenarios over the current timing plans. As stated above, the 'new plan and new TOD' combination provides the most gain for performance. The second bar shows the gains for developing new plans in combination with the old TOD intervals and still, the gains here are significant. The third bar represents the old plans implemented in combination with the new TOD's. Here, gains are still achieved, however they are only significant at certain times of the day. The new TOD selection procedure is beneficial whether significant improvements are achieved or not, in that the choice of TOD intervals can be automated and is based on historical data from all intersections, not just the critical intersection.

**Figure 63. Percent Gains for New Plans over Original Plans**

### 5.6.9 Putting It All Together

The following figures show the MOP's and gains in a more meaningful form. The costs were converted from total cost for all vehicles per day to cost per vehicle per year. This shows the gains for each vehicle achieved over the course of a year. These numbers hold much more meaning and can easily be considered for savings over a lifetime of commuting to and from work. Figure 64 shows the simulation output comparisons for the four plan scenarios to portray the yearly cost per vehicle associated with each scenario. Figure 65 shows the yearly gains per vehicle for each of the combinations of new timing plans versus the currently implemented plans. This figure confirms the significance of adopting the new plans & new TOD's, for the impact on one vehicle is impressively improved over the current performance. For instance in one year, 3.16 hours per vehicle are saved under the new timing plan with the new TOD interval.

Assuming a 30-year commute for a job, this equates to 94.8 hours saved, almost three days over the lifetime of that vehicle's commute!

**Figure 64. MOP's at 3-Intersection Corridor based on Per Vehicle Per Year**

**Figure 65. Yearly Gains of New Plans over Original Plans**

### 5.6.10 Plan Performance Over 24-Hour Period

The delay/vehicle plot over the 24-hour period that was simulated shows the periods of the day where the most improvements were achieved from the new plans. The transitional time lines for the old TOD intervals and the new TOD intervals are included in the Delay/Vehicle plots in Figure 66. The top line shows the TOD intervals for the new TOD plans, and the bottom transitional line represents the original TOD intervals currently implemented in northern Virginia. The new plan and new TOD scenario performs better than the old plan and old TOD plan at all times during the 24-hour period, however the most significant gains appear to be achieved during peak periods, i.e., the AM, MD, PM and evening periods. The new plans and the old TOD's perform similarly to the new plans and new TOD's except during the post-PM and evening periods, where the new plans and old TOD's perform much worse. This is probably due to the fact that traffic conditions most likely remain too heavy to support the off peak plan during the evening.

**Figure 66. Delay/Vehicle over 24-Hour Period at 3-Intersections**

### 5.6.11 Emissions of Timing Plans

The fuel and emission parameters control the rate at which vehicles consume fuel or emit exhaust. 'Fuel Used' was one of the MOP's evaluated in the above analysis. Here the emission parameters will be investigated to supply a performance measure of significant importance not only to commuters and traffic engineers, but also to larger concerns such as the environment. The emission values are based on the Federal Highway Administration Research and are dependant on the vehicle types, speed and acceleration/deceleration of the vehicles emitting exhaust. The three exhaust emissions in Figure 67 are carbon monoxide (CO), hydrocarbons (HC) and nitrogen-oxides ($NO_x$). The emissions produced when the old timing plans and old TOD's are implemented are the greatest, with the least emissions occurring when the new plans and new TOD's are implemented.

#### 5.6.11.1 Carbon Monoxide

Carbon monoxide (CO) is a colorless, odorless, poisonous gas (24). It is a public health problem because it enters the bloodstream through the lungs and forms carboxyhemoglobin, a compound that inhibits the blood's capacity to carry oxygen to organs and tissues. Infants, elderly persons, and individuals with respiratory diseases are also particularly sensitive. Carbon monoxide can affect healthy individuals, impairing exercise capacity, visual perception, manual dexterity, learning functions, and ability to perform complex tasks. In 1992, carbon monoxide levels exceeded the Federal air quality standard in 20 U.S. cities, home to more than 14 million people (24). Nationwide, two-thirds of the carbon monoxide emissions come from transportation sources, with the largest contribution coming from highway motor vehicles. In urban areas, the motor vehicle contribution to carbon monoxide pollution can exceed 90 percent. Carbon monoxide results from incomplete combustion of fuel and is emitted directly from vehicle tailpipes (24).

#### 5.6.11.2 HydroCarbons

Hydrocarbon emissions result when fuel molecules in the engine do not burn or burn only partially (24). Hydrocarbons react in the presence of nitrogen oxides and sunlight to form ground-level ozone, a major component of smog. Ozone irritates the eyes, damages the lungs, and aggravates respiratory problems. It is our most widespread and intractable urban air pollution problem. A number of exhaust hydrocarbons are also toxic, with the potential to cause cancer.

#### 5.6.11.3 Nitrogen Oxides

Under the high pressure and temperature conditions in an engine, nitrogen and oxygen atoms in the air react to form various nitrogen oxides, collectively known as NOx (24). Nitrogen oxides, like hydrocarbons, are precursors to the formation of ozone. Nitrogen Oxide emissions are a concern because they contribute to the formation of acid rain and, either directly or through the creation of ozone, lead to harmful effects on human health (24). According to estimates made by the U.S. Environmental Protection Agency (EPA), highway vehicles accounted for 35 percent of the 22 million tons of $NO_x$ emissions in the United States in 1995 (24).

**Figure 67. Emissions for 3-Intersection Corridor Plans**

Figure 68 shows the grams of emissions saved per day by using each of the new timing plans versus the old timing plans and old TOD's. The decrease in exhaust

emissions for improved timing plans would greatly be reduced over time. The following section compares the national emissions averages to those resulting from the 3-intersection corridor simulation.

**Figure 68. Emissions Saved for 3-Intersections Corridor over Current Plan**

### 5.6.12 Average Emissions for an "Average" Passenger Car

According to the Environmental Protection Agency, the average exhaust emissions from an “average” passenger car are listed in Table 28. These averages are compared to those resulting from the 3-intersection simulation in Figure 69. According to the 1997 averages for emissions by the EPA, all of the timing plans, even the old one, do well. There is a slight reduction, mainly in the Carbon Monoxide emissions for the new plans and new TOD’s. These figures are only approximate, due to the many assumptions imposed on both the EPA values and the simulation values and are only to be used for guidance.

Assumptions may include type of vehicle and size. Also the number of vehicles accelerating and deccelerating in the simulation may vary significantly from those making up EPA's average.

**Table 28. EPA Emissions for an "Average" Passenger Car vs. Plan Emissions**

**Figure 69. Emissions (g/mile/veh) for EPA vs. Plan Averages**

### 5.6.13 Three-Intersection Corridor Conclusions

The above case study supports the use of data mining tools for timing plan development on coordinated intersections. It has been shown that the use of system detector data for plan development versus the single-day, hand counts produce significantly improved timing plans. This method allows for much more stable volume counts since the numbers

can be taken from a historical period where variant days and traffic conditions won't be as influential on the volumes for plan development. The current method of counting cars is not reliable due to the fact that humans count cars, and for one day in which the assumption is made that traffic conditions will be "normal." It is nearly impossible to predict when traffic conditions will be "normal" or when exactly during a peak period, traffic will reach its peak. These are assumptions and educated guesses made by the traffic engineers prior to making the hand-counts. Using system detector data alleviates these issues.

The second outcome of the use of data mining tools is the production of TOD intervals based on volumes and occupancies at all intersections in the corridor. These intervals are more refined to the traffic conditions occurring throughout the corridor, whereas the current method of TOD interval selection is based primarily on aggregate, bi-directional volumes occurring at the critical intersection. For the use of the new TOD intervals with the new timing plans, only delay improved significantly when evaluated over the entire 24-hour day. However, after breaking down the 24-hour period, further analysis showed that significant improvement in the new TOD interval selection through cluster analysis did in fact provide significant improvements over certain periods of the day. These periods consisted of the majority of the newly developed TOD plan periods, supporting the success of clustering in developing meaningful timing plans and plan periods. Since there was no decline in performance across the MOP's evaluated for new plans and/or new TOD's, thus supporting the use of cluster analysis for plan development, a fully automated tool to perform this process is achievable. This tool would allow timing plans to be developed based directly on the detector data, thus

alleviating the need to choose the intervals manually through intuition, and lessening the burden on traffic engineers job by also alleviating the need for manual volume counts in the field for plan development.

# Chapter 6. Conclusions: Evaluation & Applicability

## *6.1 Research Contributions*

The major deliverables of this project are the proposed procedure (Chapter 4) and the application of data mining tools to a real-world problem. The proposed procedure directly benefits transportation engineering, while the application of cluster analysis as a basis for real-time control the systems engineering field. From the proposed procedure, the timing plan development and maintenance process can be replicated and automated. The use of data mining tools will add numerous benefits to the signal timing plan development process, especially on coordinated arterials and to the commuters using the signalized roadways. Specific benefits are summarized in the following list.

- Capability to automate timing plan development process (alleviate need for counting cars manually and avoiding issue of guessing when "normal" traffic conditions occur).
- Utilize actual data to develop more accurate timing plans and alleviating the possibility of basing the plans on a variant day.
- Utilize actual data to develop more accurate and refined TOD intervals based on all intersections and movements in corridor rather than bi-directional movements at critical intersection.
- Develop TOD intervals and timing plans based on more refined state definition, including occupancy and volume versus just volume data.
- Provide feedback for lane storage and configuration problems (turning bay lengths, the need for more lanes, etc.)
- Reduce delays, travel times, fuel used and increase the network throughput of corridors.
- Reduce harmful emissions in the environment.
- Reduce time and experience necessary by traffic engineers to develop useful timing plans and TOD intervals.
- Capability to provide up-to-date feedback of timing plan performance with ability to automate recommendations (updated timing plans): *To be investigated further in future research.*

This project contributes to systems engineering by demonstrating the use of clustering as a basis for real-time control. The use of the systems analysis process to a traffic control problem faced by traffic engineers, where the end product supports the ability to develop and maintain timing plans in real-time, is a valuable resource for systems engineering. The procedures proposed in this project provide a basis for real-time control that can be applied to many problems consisting of similar parameters; i.e., the collection of real-time data describing the state of a system as it changes through time. A valuable demonstration of systems engineering tools and methodologies to an everyday issue, resulting in a real-time, decision support system and development tool, portrays the impact and vital role of systems engineering to other fields.

This project follows a Gibson methodology (29) to reach the end product, where the problem in need of a solution is two-fold: The broad picture being the improvement of traffic movement through arterial networks and the immediate issue being the extraction of meaningful information from a large database to address the traffic movement issue. The identification of the problem at hand resulted in goals and alternatives to address the data and traffic issues. The goals set forth here were to implement data mining tools that would organize the information contained in the data to support traffic control tools for improved efficiency. The alternatives to reach the goal included alternate data mining tools and algorithms. The format of the volume and occupancy data and the scope of he current means of traffic control resulted in the selection of the best alternative solution: hierarchical, centroid cluster analysis. This alternative was selected based on clustering research studies and analysis on sample data sets. The selection of this solution for data organization was tested with cluster validation techniques to ensure the solution of cluster

analysis for meaningful data interpretation was valid. Cluster validation was a vital step in the systems analysis approach to solving a problem. Often a solution is proposed and untested under the assumption the solution is valid and optimal (29). This project includes the systems analysis approach, which showed that hierarchical cluster analysis does create meaningful clusters from the data and so the information extracted from the database can be implemented as a valid tool for signal development.

The result of a thorough systems approach to problem solving is shown here to improve traffic performance through corridors with the ability to support automated tools for plan development and maintenance. The underlying application of data mining tools to traffic data in this project can be used as a guide for the development of similar real-time, automated control tools. This demonstration of systems engineering tools and analysis to a real-world problem is a valuable contribution to the systems engineering field.

### *6.2 Usability of Procedure*

The usefulness and usability of the proposed procedure must be considered for effectiveness. The primary user group considered here is the VDOT northern Virginia traffic signal control group. It was this center that supplied the data for this research and who benefit first-hand from these research investigations. The signal control method to improve upon introduced here is the time-of-day (TOD) method signal control. Since this is the primary technique supported throughout the country, this procedure could be adapted and utilized by any DOT that retrieves system detector data. System detectors

are becoming increasingly widespread and the full capability of utilizing such data has not yet been fully realized, especially for the basic forms of signal control such as TOD.

The procedure allows for fully automated plan development and maintenance tools; however, further research should be continued to reach this stage. The proposed procedure here supports the use of detector data for improved timing plan development at single intersections, where it has been shown that performance improves significantly when timing plans are developed from historical data. This project has also supported the claim that corridors do experience an increase in performance with the clustered TOD intervals; however, these improvements are not as significant as those experienced by the data base timing plan development. Even in the case that there would be no change in the performance of the system with the newly developed timing plan and TOD intervals, since it would not degrade the performance, the procedure could still be utilized for automation of the signal development and maintenance procedure. The procedure is useful for all cases in that it allows for TOD intervals to be selected automatically based on traffic data versus engineers' intuitions for corridors. These intervals are also based on individual volume and occupancy values collected continuously over the period of multiple days to better represent actual traffic conditions. Finally, this proposed procedure introduces a method of data mining for maintenance and feedback of signal performance over time to better alert the need for change and updating timing plans. However, this area was not fully supported due to the scope of this research and should be investigated in the future.

This procedure could benefit many traffic engineers that deal with traffic signal control through signalized intersections. It not only speeds up the timing plan

development process and TOD interval selection process, basing it on historical data, but it also allows for improved timing plans to be developed at single intersections for improved performance, without the need to manually count cars.

### *6.3 Simulation as Realistic Representation*

The simulation outputs presented here are the main support of the effectiveness of the proposed procedure, so it is important to consider how accurately these results represent actual traffic conditions. SimTraffic accounts for conditions such as driver behavior characteristics (aggressive, etc.), types of vehicles (trucks versus cars), road type and grade, etc. Conditions that would not be accountable for would be things such as weather and incidents. But these types of conditions may exist in the data used to drive the simulation. This presents the ability to model traffic conditions as accurately as possible in that actual 15-minute volumes can be fed into the simulation for recording the outputs of the timing plan effectiveness for such traffic conditions. The simulations can be run repetitively using alternate random number seeds for representing a dynamic simulation. SimTraffic is also capable of simulating during periods of transition between timing plans to account for transition effects on performance. Overall, the simulation is a fairly accurate display of actual traffic conditions, especially with the detailed input available for setting up the roadways and driver characteristics (16). The simulation may be a better representation if occupancies could also be considered in the simulation as well as volumes, but since these values are correlated it is not essential. Calibration of the simulation tool using actual counts collected at the intersections would also provide a better-supported representation of actual traffic conditions.

### *6.4 Future Research*

Due to the exploratory nature of this project, there remains much research to be done before the optimal tool can be created for timing plan development and maintenance. A procedure has been proposed for the enhancement of signal timing plans through the use of system detector data. The purpose here is to show that detector data can be utilized to simplify the timing plan development process, as well as to allow a means for constant feedback on timing plans' performance. The use of detector data also allows for the timing plans to be better prepared to handle actual traffic conditions based on the historical data base of traffic trends. The areas that need further investigation are as follows:

- The effects of increased transitions on corridor performance
- Detailed analysis supporting the optimal clustering methodology
- Identification and performance of reduced state space (Select intersections, detector, etc.)
- Appropriate time to extract data from the database for timing plan development (Accounting for variance in traffic conditions over time)
- Weighting the cluster input variables for optimal results (Importance of intersections, detectors, etc.)
- Appropriate simulation tool for monitoring timing plan performance
- Detailed analysis of simulation performance outputs (Queue length, v/c ratio, MOP's at individual intersections, etc.)
- Verification of detector data and turning lane conversion factors with the SmartTravelVan
- Classification as a tool for plan maintenance

- Investigation of replication criteria for determination of the appropriate number of clusters
- Investigate lift achieved from hand-picking a greater number of TOD intervals for comparison of lift achieved by cluster analysis

### 6.4.1 Cluster Methodology Analysis

The cluster methodologies were only briefly investigated to propose a sufficient method for clustering volume and occupancy data. The centroid cluster methodology was proposed here, however an in-depth investigation into all of the possible methodologies would be appropriate before proposing a finalized procedure. Each methodology should be evaluated based on dissimilarity metrics and performance on different sized data sets. The performance of these cluster methodologies should be evaluated using internal cluster validation. The evaluation for this procedure is brief, but sufficient for exploratory research.

### 6.4.2 Transition Effects on Corridor Performance

Transitioning between timing plans is an extremely complex event for signal plan development. There are alternate methods for a controller to transition between plans and much research has been done on the best method of transition. The simulation package used for this research only transitions from plan to plan in one way that simTraffic utilizes as discussed in Chapter 2. However, there are multiple ways for a transition to occur and so simulation packages that can compare these alternate methods should be investigated. Also, the procedure introduced here typically constructs more plans per day, often transitioning from an existing plan to another and back. This increased

transition may in fact impede the performance of traffic control due to the setbacks that may arise in traffic flow during transitional times. Brief transition effects can be observed from comparing the outputs of the performance of the same timing plans operating under the original TOD intervals and the clustered TOD intervals. Since there were more TOD intervals created by clustering, the outputs of this simulation may provide some insight as to how severe the increased number of transitions are to performance. However, it is essential that the effect of transitions be thoroughly investigated, with a suggestion for the number of transitions acceptable before performance deteriorates.

### 6.4.3 Reduced State Space

A major advantage of using detector data and data mining tools to develop timing plans and TOD intervals is that the rawest form of the data can be used, i.e., volume and occupancy at each system detector for each intersection from all historical data. It may not be necessary to use such a detailed state space in developing clusters. It should be determined what the optimal state space is with the least amount of variables involved. This will simplify the state space, which in turn will speed up the analysis. For instance, small intersections may not contribute to the clusters or timing plans and thus could be excluded from the analysis without any loss of TOD interval and plan resolution. In this research a brief analysis was done on cluster development with the use of only volume values versus volume and occupancy values. It was found that using both volume and occupancy provided better-tailored plans for traffic conditions; however, there may exist intersections and/or phases that do not contribute significantly to the analysis and should

be removed. Classification would be one form of determining variable importance once the clusters have been developed.

### 6.4.4 Historical Data Period

The period of time for which historical data is extracted from the database for cluster analysis and timing plan development may also be crucial to accurate results. There is a point when going too far back in time to collect detector data will deteriorate the results. With years of data available, it may be tempting to construct an overly detailed analysis. Traffic varies over time and an approximate historical cut-off for data collection should be determined such that the results are not influenced by out-dated traffic trends. An investigation into the variance in traffic over time and the significance of these changes for timing plans is recommended. This analysis should take into account seasonal trends and the development of alternate plans for different seasons where traffic conditions vary significantly. The forecasting of traffic conditions based on historical trends can also be investigated to further enhance the plan maintenance portion of this research.

### 6.4.5 Weighting of Cluster Input Variables

An extremely important consideration is the weighting of input variables for cluster analysis. This information can be determined with the effectiveness of a reduced state space. A process such as classification will provide variable importance scores, which can be used to determine appropriate weights as the input variables contribute to the cluster analysis. For instance, it may be beneficial to weight the detectors at the critical intersection more heavily than those at a much smaller intersection so the clusters will be

affected by the more important intersection in the corridor. The volumes and occupancies may also need to be weighted differently. This result was shown briefly in the analysis of the un-standardized volumes and occupancies and the single case in which standardized volumes were weighted more heavily than the standardized occupancies. Both of these methods weight volume over occupancy as is natural from the raw data since occupancy values are much smaller than volumes. The brief analysis done here shows that in fact the clusters are cleaner and more appropriate when more weight is given to the volume, although it should be determined if this is a significant improvement or not.

### 6.4.6 Simulation Tool

SimTraffic was used to run simulations for this research primarily because that is the tool used by VDOT. Since the data here has been provided by VDOT and this research is to benefit the work done at the northern Virginia traffic signal control center, the procedure introduced here is catering to a specific user and the timing plan development process should mimic the current means. However, timing plan development in Synchro can be used in correlation with other simulation tools such as Corsim and Transyt 7-F. SimTraffic provided the ability to simulate over alternate timing plans to address transition times as well as the ability to simulate based on an off-line 15-minute volume file straight from the database for simulation parameters. This allows for realistic results based on actual traffic conditions. Another benefit to SimTraffic is that the number of intersection allowable by the software is extremely large (> 100) and so intricate networks can me modeled. SimTraffic only allows 19 intervals (15-minute intervals) to

be simulated at a time, resulting in a tedious process. There has also been the release of the newest version of SimTraffic, version 5.0, which may provide further advantages for simulating timing plans. The strengths and weaknesses of Corsim and Transyt 7-F were not thoroughly investigated and may enhance the simulation process for future analyses. This may be an important consideration for further analysis into timing plan performance.

### 6.4.7 Simulation Outputs (MOP's)

The measures of performance (MOP's) provided by SimTraffic are numerous. The analysis in this project only examines outputs for the entire corridor and for the full day. Future analysis should include comparisons of plan performance during different times of the day and at individual intersections. This will allow for the critical time intervals and intersections to be identified to better understand where the timing plans are the weakest. Results can also be compared on a movement or phase basis, looking into which direction traffic suffers at different times of the day under the alternate timing plans. There are also numerous outputs from SimTraffic not used in this project such as queue lengths and timing plan measures of effectiveness. There are many forms to display and dissect the simulation outputs, which may provide further insight into the conditions experienced under alternate timing plans.

### 6.4.8 Verification of Detector Data with the SmartTravelVan

The Smart Travel Laboratory and Virginia Research Council own a SmartTravelVan, which acts as a mobile video detection system. This can be placed at any intersection for traffic data collection. The accuracy of the detector data can be evaluated with the

SmartTravelVan data results. Also, since no system detectors exist on many of the turning lane movements, the SmartTravelVan can be used to collect this data for more appropriate turning movement data. This will greatly aid in the cluster development and simulation of the timing plans, since the turning movements where system detectors do not exist must be approximated based on data collected by VDOT. These collections can also be used to calibrate the simulation parameters for accurate timing plans and simulations.

### 6.4.9 Classification as a tool for plan maintenance

CART was introduced in Chapter 3 as a tool for cluster validation. The analysis from such an output can be taken one step further to provide a feedback mechanism for traffic engineers. The trees produced by the classification of the clustered data provide splitting rules for which proved to perform well ( > 90% for both single intersections and corridors). This would allow automation of the performance of the newly clustered and implemented timing plans by monitoring each current 15-minute traffic state that occurs and classifying it into the correct cluster or timing plan. This procedure would not have to be used with clustered data, but could be used on the current method, where the assigned timing plan becomes the target variable by which the classification rule is formed from the corresponding input variables. This notion extends the idea of the use of data mining tools for the enhancement of traffic signal maintenance and would alert traffic engineers of the need to update timing plans or adjust the TOD intervals. This is a major component of automated timing plan tools and should be investigated as to its ability to correctly identify outdated plans.

### 6.4.10 Investigation of replication criteria

Recent studies by Atlas and Overall (14) have investigated stopping rules that uncover cluster levels appropriate for overlapping clusters. This may be useful for the traffic data, since the volume and occupancy pairs over a 24-hour period do not form completely distinct groups prior to clustering. It is recommended that studies be conducted with the replication criteria for higher-order clustering on split-sample means and compared to the results of the pseudo F, T2 and CCC stopping rules, which have shown in this research to perform well for determination of TOD intervals and timing plans.

### 6.4.11 Hand-pick increased number of TOD Intervals

An interesting study would be to handpick the TOD intervals and plan transition times for a 24-hour period based on the number of clusters recommended (7 TOD intervals in the 3-intersection case study). These intuition-based TOD intervals could be simulated and the performance results achieved from these handpicked intervals compared to those resulting from the cluster analysis. The level of improvement in MOP's can be analyzed by the use of clustered TOD intervals for 7 clusters versus handpicking 7 TOD intervals. It is possible that a significant amount of performance improvement results from the addition of the number of plans implemented and an analysis should be conducted on the value of the clustered selection of plans and intervals. However, the cluster analysis would still be valuable to determine the number of intervals and plans that should be selected based on traffic data.

## *6.5 Research Discoveries*

Discoveries have been made in lieu of the research done here that were not initially anticipated. One find is that clusters can alert traffic engineers of critical conditions that arise in certain situations. For instance, if enough saturation is experienced by any movement in an intersection, the occupancies would likely form a cluster for that situation, which my not necessarily be taken care of with an additional timing plan due to the severity of the situation. This would alert the traffic engineers of the need for extended or additional turning bays or alternate lane configurations to better support existing traffic demand.

Another important discovery is that the input variables for the cluster analysis should weight volume heavier than occupancy. The variables should also be standardized to produce the cleanest TOD intervals from the clusters. The two recommended formats for cluster variables is either standardized volumes and occupancies, with volumes weight by a factor of 20, or standardized volumes and occupancies with occupancies reduced to values of $< 26$. Of course for insight into the need to alter lane configurations or saturated movements problems, the occupancies should be left untouched. However, for plan development, it is necessary to alleviate the formation of random clusters not represented by a particular time-of-day.

The project also recommends from the sensitivity studies, that a minimum number of observations should exist in each cluster and should be applied to the cluster algorithm. This produces substantial clusters that exist for sufficient times, in order to be supported by an entire timing plan. Otherwise, clusters may be developed based on one or two erroneous observations that cannot be supported by a timing plan and thus take away from the refinement of the remaining clusters during the 24-hour period.

Finally, it was discovered that TOD interval selection does become increasingly difficult in a corridor versus a single intersection. Thus the cluster procedure is highly recommended, as corridors become increasingly large and complex for the identification of TOD intervals. Single intersections are easily defined with appropriate TOD intervals by traffic engineers, however the single-day, hand-counts at single intersections and corridors is an out-dated procedure and should be replaced by the use of historical data for plan development.

## 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. Tarnoff, Philip. Traffic Signal Control class notes. February 7, 2000, Northern Virginia Traffic Control Center, Fairfax, Virginia.

6. Haenel, Herman. *Research Initiatives for Traffic Signal Control Systems*. Transportation Research Circular, Number 380, October 1991, ISSN 0097-8515.

7. Brydia, Eisele and Turner. Development of an ITS Data Management System. Texas A&M University prepared for TRB, January 1998.

8. Remer, Klugman, Hildebrand and Belrose. *Addition of Adaptive Control to the Minneapolis Signal System: Phase 2 Issues and Solutions*. Federal Highway Administration and the City of Minneapolis prepared for TRB, January 1998.

9. 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.

10. *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 1*. SAS Institute Inc. Cary, NC, 1990.

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

12. Jain, Dubes. *Algorithms for Clustering Data*. Prentice Hall, Inc., New Jersey, 1988.

13. Kaufman, Rousseeuw. *Finding Groups in Data – An Introduction to Cluster Analysis*. Wiley-Interscience, John Wiley & Sons, Inc., New York, 1990.

14. Atlas, Overall. *Comparative Evaluation of Two Superior Stopping Rules for Hierarchical Cluster analysis.* Psychometrika, Vol. 59, No. 4, pp. 581 – 591. December 1994.

15. *Adaptation from SAS / STAT User's Guide (1990) and Sarle and Kuo (1993).* Copyright 1996 by SAS Institute Inc, Cary, NC, USA.

16. Husch, Albeck. *Synchro Plus 4.0 User' Guide*. Trafficware, Albany, California, 1999.

17. Lapionte, F. and Legendre, P. *A Classification of Pure Malt Scotch Whiskies*. Applied Statistics, Vol. 43, No. 1, 1994, pp. 237 – 257.

18. May, A. *Traffic Flow Fundamentals*. Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1990.

19. Glymour, Madigan, Pregibon, Smyth. *Statistical Themes and Lessons for Data Mining*. Kluwer Academic Publishers. Data Mining and Knowledge Discovery, 1, 11 – 28, 1997.

20. Data Mining: Extending the Information Warehouse Framework. http://www.almaden.ibm.com/cs/quest/papers/whitepaper.html.

21. Seal. *Multivariate Statistical Analysis for Biologists*. Spottiswoode, Ballantyne & Co., Ltd., London, 1966.

22. Everitt. *Cluster Analysis, Third Edition*. Halsted Press, an imprint of John Wiley & Sons, New York, 1993.

23. Jolliffe, I.T. *Principal Components Analysis*. Springer-Verlag, New York, 1986.

24. Environmental Protection Agency, Volume 2, Issue 5.3, April 2001. http://www.epa.gov/.

25. Jain, Dubes. *Validity Studies in Clustering Methodologies*. Pattern Recognition, Vol. 11, pp. 234 – 254. Pergamom Press, Great Britain, 1979.

26. Averill M. Law & Associates. *ExpertFit User's Guide*. January 2001 Edition.

27. SAS Institute Inc., SAS Technical Report: A-108. *Cubic Clustering Criterion*, Cary, NC: SAS Institute Inc., 1983. 56 pp.

28. Devore. *Probability and Statistics for Engineering and the Sciences, Fourth Edition*. Brooks/Cole Publishing Co., U.S.A., 1995.

29. Gibson, John. *How To Do a Systems Analysis and System Analyst Decalog*. UVA Printing & Copying Services, 1999.

# Appendix A – 3-Intersection Corridor CPCC Matrices for 7 Clusters

## *Cluster 1*

| Variable | V2SH | O2SH | V4SH | O4SH | V6SH | O6SH | V8SH | O8SH | V2BLMT |
|---|---|---|---|---|---|---|---|---|---|
| V2SH | 1.00000 | 0.94429 | 0.45032 | 0.40415 | 0.29949 | 0.16820 | 0.50753 | 0.85743 | 0.80473 |
| V2SH | | <.0001 | 0.1418 | 0.1926 | 0.3443 | 0.6013 | 0.0921 | 0.0004 | 0.0016 |
| O2SH | 0.94429 | 1.00000 | 0.23673 | 0.16834 | 0.06725 | 0.10796 | 0.30694 | 0.69348 | 0.60042 |
| O2SH | <.0001 | | 0.4588 | 0.6010 | 0.8355 | 0.7384 | 0.3318 | 0.0124 | 0.0390 |
| V4SH | 0.45032 | 0.23673 | 1.00000 | 0.90236 | 0.85009 | 0.56188 | 0.86845 | 0.68753 | 0.79065 |
| V4SH | 0.1418 | 0.4588 | | <.0001 | 0.0005 | 0.0573 | 0.0002 | 0.0135 | 0.0022 |
| O4SH | 0.40415 | 0.16834 | 0.90236 | 1.00000 | 0.70986 | 0.32855 | 0.71550 | 0.59457 | 0.69631 |
| O4SH | 0.1926 | 0.6010 | <.0001 | | 0.0097 | 0.2971 | 0.0089 | 0.0414 | 0.0119 |
| V6SH | 0.29949 | 0.06725 | 0.85009 | 0.70986 | 1.00000 | 0.66906 | 0.91093 | 0.59039 | 0.77320 |
| V6SH | 0.3443 | 0.8355 | 0.0005 | 0.0097 | | 0.0173 | <.0001 | 0.0433 | 0.0032 |
| O6SH | 0.16820 | 0.10796 | 0.56188 | 0.32855 | 0.66906 | 1.00000 | 0.59400 | 0.30114 | 0.41781 |
| O6SH | 0.6013 | 0.7384 | 0.0573 | 0.2971 | 0.0173 | | 0.0417 | 0.3415 | 0.1765 |
| V8SH | 0.50753 | 0.30694 | 0.86845 | 0.71550 | 0.91093 | 0.59400 | 1.00000 | 0.79119 | 0.88743 |
| V8SH | 0.0921 | 0.3318 | 0.0002 | 0.0089 | <.0001 | 0.0417 | | 0.0022 | 0.0001 |
| O8SH | 0.85743 | 0.69348 | 0.68753 | 0.59457 | 0.59039 | 0.30114 | 0.79119 | 1.00000 | 0.93568 |
| O8SH | 0.0004 | 0.0124 | 0.0135 | 0.0414 | 0.0433 | 0.3415 | 0.0022 | | <.0001 |
| V2BLMT | 0.80473 | 0.60042 | 0.79065 | 0.69631 | 0.77320 | 0.41781 | 0.88743 | 0.93568 | 1.00000 |
| V2BLMT | 0.0016 | 0.0390 | 0.0022 | 0.0119 | 0.0032 | 0.1765 | 0.0001 | <.0001 | |
| O2BLMT | 0.84697 | 0.69694 | 0.65457 | 0.56496 | 0.60576 | 0.36737 | 0.77726 | 0.89418 | 0.92471 |
| O2BLMT | 0.0005 | 0.0118 | 0.0209 | 0.0556 | 0.0368 | 0.2401 | 0.0029 | <.0001 | <.0001 |
| V4BLMT | 0.68858 | 0.79573 | -0.24122 | -0.15048 | -0.45212 | -0.27868 | -0.23872 | 0.32457 | 0.16289 |
| V4BLMT | 0.0133 | 0.0020 | 0.4501 | 0.6406 | 0.1400 | 0.3804 | 0.4549 | 0.3033 | 0.6130 |
| O4BLMT | 0.52219 | 0.62108 | -0.28019 | -0.10973 | -0.53137 | -0.45362 | -0.41009 | 0.12955 | 0.00680 |
| O4BLMT | 0.0816 | 0.0311 | 0.3777 | 0.7343 | 0.0754 | 0.1386 | 0.1855 | 0.6882 | 0.9833 |
| V6BLMT | 0.92887 | 0.80152 | 0.65871 | 0.60443 | 0.55069 | 0.33843 | 0.72695 | 0.95077 | 0.92500 |
| V6BLMT | <.0001 | 0.0017 | 0.0198 | 0.0374 | 0.0635 | 0.2819 | 0.0074 | <.0001 | <.0001 |
| O6BLMT | 0.96846 | 0.89597 | 0.51845 | 0.43084 | 0.38151 | 0.30146 | 0.60467 | 0.92496 | 0.84107 |
| O6BLMT | <.0001 | <.0001 | 0.0842 | 0.1620 | 0.2211 | 0.3410 | 0.0373 | <.0001 | 0.0006 |
| V6ND | 0.96439 | 0.85389 | 0.57027 | 0.53939 | 0.44868 | 0.28061 | 0.61927 | 0.90962 | 0.88339 |
| V6ND | <.0001 | 0.0004 | 0.0529 | 0.0703 | 0.1435 | 0.3770 | 0.0318 | <.0001 | 0.0001 |

Cluster 1 (Con't.)

| Variable | O2BLMT | V4BLMT | O4BLMT | V6BLMT | O6BLMT | V6ND | O6ND | V8ND | O8ND |
|---|---|---|---|---|---|---|---|---|---|
| V2SH | 0.84697 | 0.68858 | 0.52219 | 0.92887 | 0.96846 | 0.96439 | 0.97898 | 0.82189 | 0.83487 |
| V2SH | 0.0005 | 0.0133 | 0.0816 | <.0001 | <.0001 | <.0001 | <.0001 | 0.0010 | 0.0007 |
| O2SH | 0.69694 | 0.79573 | 0.62108 | 0.80152 | 0.89597 | 0.85389 | 0.93242 | 0.88202 | 0.88946 |
| O2SH | 0.0118 | 0.0020 | 0.0311 | 0.0017 | <.0001 | 0.0004 | <.0001 | 0.0001 | 0.0001 |
| V4SH | 0.65457 | -0.24122 | -0.28019 | 0.65871 | 0.51845 | 0.57027 | 0.43197 | -0.03405 | -0.01156 |
| V4SH | 0.0209 | 0.4501 | 0.3777 | 0.0198 | 0.0842 | 0.0529 | 0.1608 | 0.9163 | 0.9716 |
| O4SH | 0.56496 | -0.15048 | -0.10973 | 0.60443 | 0.43084 | 0.53939 | 0.33595 | -0.03313 | -0.01930 |
| O4SH | 0.0556 | 0.6406 | 0.7343 | 0.0374 | 0.1620 | 0.0703 | 0.2857 | 0.9186 | 0.9525 |
| V6SH | 0.60576 | -0.45212 | -0.53137 | 0.55069 | 0.38151 | 0.44868 | 0.27863 | -0.24443 | -0.21581 |
| V6SH | 0.0368 | 0.1400 | 0.0754 | 0.0635 | 0.2211 | 0.1435 | 0.3805 | 0.4439 | 0.5005 |
| O6SH | 0.36737 | -0.27868 | -0.45362 | 0.33843 | 0.30146 | 0.28061 | 0.23173 | -0.09761 | -0.06992 |
| O6SH | 0.2401 | 0.3804 | 0.1386 | 0.2819 | 0.3410 | 0.3770 | 0.4686 | 0.7628 | 0.8290 |
| V8SH | 0.77726 | -0.23872 | -0.41009 | 0.72695 | 0.60467 | 0.61927 | 0.48038 | -0.02564 | 0.00436 |
| V8SH | 0.0029 | 0.4549 | 0.1855 | 0.0074 | 0.0373 | 0.0318 | 0.1139 | 0.9370 | 0.9893 |
| O8SH | 0.89418 | 0.32457 | 0.12955 | 0.95077 | 0.92496 | 0.90962 | 0.86239 | 0.53214 | 0.55479 |
| O8SH | <.0001 | 0.3033 | 0.6882 | <.0001 | <.0001 | <.0001 | 0.0003 | 0.0749 | 0.0612 |
| V2BLMT | 0.92471 | 0.16289 | 0.00680 | 0.92500 | 0.84107 | 0.88339 | 0.77088 | 0.36494 | 0.39080 |
| V2BLMT | <.0001 | 0.6130 | 0.9833 | <.0001 | 0.0006 | 0.0001 | 0.0033 | 0.2435 | 0.2091 |
| O2BLMT | 1.00000 | 0.30318 | 0.09879 | 0.87207 | 0.85466 | 0.86626 | 0.81296 | 0.51597 | 0.53983 |
| O2BLMT | | 0.3381 | 0.7600 | 0.0002 | 0.0004 | 0.0003 | 0.0013 | 0.0859 | 0.0700 |
| V4BLMT | 0.30318 | 1.00000 | 0.92354 | 0.45597 | 0.60172 | 0.57753 | 0.68390 | 0.94263 | 0.93428 |
| V4BLMT | 0.3381 | | <.0001 | 0.1363 | 0.0385 | 0.0492 | 0.0142 | <.0001 | <.0001 |
| O4BLMT | 0.09879 | 0.92354 | 1.00000 | 0.29412 | 0.38838 | 0.42742 | 0.49134 | 0.79844 | 0.78192 |
| O4BLMT | 0.7600 | <.0001 | | 0.3534 | 0.2122 | 0.1658 | 0.1048 | 0.0018 | 0.0027 |
| V6BLMT | 0.87207 | 0.45597 | 0.29412 | 1.00000 | 0.96310 | 0.98407 | 0.90940 | 0.61013 | 0.63065 |
| V6BLMT | 0.0002 | 0.1363 | 0.3534 | | <.0001 | <.0001 | <.0001 | 0.0351 | 0.0279 |
| O6BLMT | 0.85466 | 0.60172 | 0.38838 | 0.96310 | 1.00000 | 0.97002 | 0.97868 | 0.76760 | 0.78565 |
| O6BLMT | 0.0004 | 0.0385 | 0.2122 | <.0001 | | <.0001 | <.0001 | 0.0036 | 0.0025 |
| V6ND | 0.86626 | 0.57753 | 0.42742 | 0.98407 | 0.97002 | 1.00000 | 0.94257 | 0.71450 | 0.73226 |
| V6ND | 0.0003 | 0.0492 | 0.1658 | <.0001 | <.0001 | | <.0001 | 0.0090 | 0.0068 |

| Variable | V2SH | O2SH | V4SH | O4SH | V6SH | O6SH | V8SH | O8SH | V2BLMT |
|---|---|---|---|---|---|---|---|---|---|
| O6ND | 0.97898 | 0.93242 | 0.43197 | 0.33595 | 0.27863 | 0.23173 | 0.48038 | 0.86239 | 0.77088 |
| O6ND | <.0001 | <.0001 | 0.1608 | 0.2857 | 0.3805 | 0.4686 | 0.1139 | 0.0003 | 0.0033 |
| V8ND | 0.82189 | 0.88202 | -0.03405 | -0.03313 | -0.24443 | -0.09761 | -0.02564 | 0.53214 | 0.36494 |
| V8ND | 0.0010 | 0.0001 | 0.9163 | 0.9186 | 0.4439 | 0.7628 | 0.9370 | 0.0749 | 0.2435 |
| O8ND | 0.83487 | 0.88946 | -0.01156 | -0.01930 | -0.21581 | -0.06992 | 0.00436 | 0.55479 | 0.39080 |
| O8ND | 0.0007 | 0.0001 | 0.9716 | 0.9525 | 0.5005 | 0.8290 | 0.9893 | 0.0612 | 0.2091 |

CLUSTER = 1

| Variable | O2BLMT | V4BLMT | O4BLMT | V6BLMT | O6BLMT | V6ND | O6ND | V8ND | O8ND |
|---|---|---|---|---|---|---|---|---|---|
| O6ND | 0.81296 | 0.68390 | 0.49134 | 0.90940 | 0.97868 | 0.94257 | 1.00000 | 0.85098 | 0.86509 |
| O6ND | 0.0013 | 0.0142 | 0.1048 | <.0001 | <.0001 | <.0001 | | 0.0004 | 0.0003 |
| V8ND | 0.51597 | 0.94263 | 0.79844 | 0.61013 | 0.76760 | 0.71450 | 0.85098 | 1.00000 | 0.99938 |
| V8ND | 0.0859 | <.0001 | 0.0018 | 0.0351 | 0.0036 | 0.0090 | 0.0004 | | <.0001 |
| O8ND | 0.53983 | 0.93428 | 0.78192 | 0.63065 | 0.78565 | 0.73226 | 0.86509 | 0.99938 | 1.00000 |
| O8ND | 0.0700 | <.0001 | 0.0027 | 0.0279 | 0.0025 | 0.0068 | 0.0003 | <.0001 | |

***Cluster 2***

| Variable | V2SH | O2SH | V4SH | O4SH | V6SH | O6SH | V8SH | O8SH | V2BLMT |
|---|---|---|---|---|---|---|---|---|---|
| V2SH | 1.00000 | 0.93197 | 0.28925 | 0.38098 | 0.34252 | 0.72643 | 0.53464 | 0.48015 | 0.82883 |
| V2SH | | <.0001 | 0.1434 | 0.0499 | 0.0803 | <.0001 | 0.0041 | 0.0113 | <.0001 |
| O2SH | 0.93197 | 1.00000 | 0.37868 | 0.49141 | 0.42879 | 0.66589 | 0.65857 | 0.59053 | 0.79588 |
| O2SH | <.0001 | | 0.0514 | 0.0092 | 0.0256 | 0.0001 | 0.0002 | 0.0012 | <.0001 |
| V4SH | 0.28925 | 0.37868 | 1.00000 | 0.93835 | 0.29516 | 0.57569 | 0.28893 | -0.09756 | -0.09764 |
| V4SH | 0.1434 | 0.0514 | | <.0001 | 0.1350 | 0.0017 | 0.1438 | 0.6283 | 0.6280 |
| O4SH | 0.38098 | 0.49141 | 0.93835 | 1.00000 | 0.24214 | 0.57857 | 0.27479 | -0.06449 | -0.01077 |
| O4SH | 0.0499 | 0.0092 | <.0001 | | 0.2237 | 0.0016 | 0.1654 | 0.7493 | 0.9575 |
| V6SH | 0.34252 | 0.42879 | 0.29516 | 0.24214 | 1.00000 | 0.62704 | 0.69326 | 0.42707 | 0.30818 |
| V6SH | 0.0803 | 0.0256 | 0.1350 | 0.2237 | | 0.0005 | <.0001 | 0.0263 | 0.1178 |
| O6SH | 0.72643 | 0.66589 | 0.57569 | 0.57857 | 0.62704 | 1.00000 | 0.53805 | 0.20965 | 0.43562 |
| O6SH | <.0001 | 0.0001 | 0.0017 | 0.0016 | 0.0005 | | 0.0038 | 0.2939 | 0.0231 |
| V8SH | 0.53464 | 0.65857 | 0.28893 | 0.27479 | 0.69326 | 0.53805 | 1.00000 | 0.87755 | 0.66686 |
| V8SH | 0.0041 | 0.0002 | 0.1438 | 0.1654 | <.0001 | 0.0038 | | <.0001 | 0.0001 |
| O8SH | 0.48015 | 0.59053 | -0.09756 | -0.06449 | 0.42707 | 0.20965 | 0.87755 | 1.00000 | 0.78094 |
| O8SH | 0.0113 | 0.0012 | 0.6283 | 0.7493 | 0.0263 | 0.2939 | <.0001 | | <.0001 |
| V2BLMT | 0.82883 | 0.79588 | -0.09764 | -0.01077 | 0.30818 | 0.43562 | 0.66686 | 0.78094 | 1.00000 |
| V2BLMT | <.0001 | <.0001 | 0.6280 | 0.9575 | 0.1178 | 0.0231 | 0.0001 | <.0001 | |
| O2BLMT | 0.70031 | 0.79871 | 0.04868 | 0.19504 | 0.50678 | 0.42819 | 0.68634 | 0.71289 | 0.80548 |
| O2BLMT | <.0001 | <.0001 | 0.8095 | 0.3276 | 0.0070 | 0.0259 | <.0001 | <.0001 | <.0001 |
| V4BLMT | 0.27464 | 0.35791 | 0.38968 | 0.51712 | -0.18727 | 0.08669 | -0.02391 | -0.06367 | 0.07781 |
| V4BLMT | 0.1656 | 0.0668 | 0.0445 | 0.0057 | 0.3496 | 0.6672 | 0.9058 | 0.7524 | 0.6997 |
| O4BLMT | -0.23168 | -0.03114 | 0.10176 | 0.15117 | -0.19285 | -0.30450 | 0.09221 | 0.13231 | -0.07329 |
| O4BLMT | 0.2449 | 0.8775 | 0.6135 | 0.4516 | 0.3352 | 0.1225 | 0.6474 | 0.5106 | 0.7164 |
| V6BLMT | -0.24981 | -0.05411 | 0.03180 | 0.00563 | 0.53590 | -0.15439 | 0.24697 | 0.16997 | -0.18835 |
| V6BLMT | 0.2089 | 0.7887 | 0.8749 | 0.9778 | 0.0040 | 0.4419 | 0.2143 | 0.3967 | 0.3468 |
| O6BLMT | -0.40310 | -0.23262 | 0.28865 | 0.27480 | 0.27091 | -0.18170 | -0.03545 | -0.18556 | -0.49274 |
| O6BLMT | 0.0371 | 0.2430 | 0.1442 | 0.1654 | 0.1717 | 0.3644 | 0.8607 | 0.3541 | 0.0090 |
| V6ND | -0.19325 | -0.01522 | 0.28035 | 0.26008 | 0.69460 | 0.15060 | 0.25708 | 0.00807 | -0.28407 |
| V6ND | 0.3342 | 0.9400 | 0.1567 | 0.1901 | <.0001 | 0.4534 | 0.1955 | 0.9681 | 0.1510 |

Cluster 2 (Con't.)

| Variable | O2BLMT | V4BLMT | O4BLMT | V6BLMT | O6BLMT | V6ND | O6ND | V8ND | O8ND |
|---|---|---|---|---|---|---|---|---|---|
| V2SH | 0.70031 | 0.27464 | -0.23168 | -0.24981 | -0.40310 | -0.19325 | -0.12833 | 0.67514 | 0.65776 |
| V2SH | <.0001 | 0.1656 | 0.2449 | 0.2089 | 0.0371 | 0.3342 | 0.5235 | 0.0001 | 0.0002 |
| O2SH | 0.79871 | 0.35791 | -0.03114 | -0.05411 | -0.23262 | -0.01522 | 0.05510 | 0.69602 | 0.67795 |
| O2SH | <.0001 | 0.0668 | 0.8775 | 0.7887 | 0.2430 | 0.9400 | 0.7849 | <.0001 | 0.0001 |
| V4SH | 0.04868 | 0.38968 | 0.10176 | 0.03180 | 0.28865 | 0.28035 | -0.16629 | -0.07609 | 0.09404 |
| V4SH | 0.8095 | 0.0445 | 0.6135 | 0.8749 | 0.1442 | 0.1567 | 0.4071 | 0.7060 | 0.6408 |
| O4SH | 0.19584 | 0.51712 | 0.15117 | 0.00563 | 0.27480 | 0.26008 | -0.23806 | 0.07822 | 0.26382 |
| O4SH | 0.3276 | 0.0057 | 0.4516 | 0.9778 | 0.1654 | 0.1901 | 0.2318 | 0.6982 | 0.1836 |
| V6SH | 0.50678 | -0.18727 | -0.19285 | 0.53590 | 0.27091 | 0.69460 | 0.65370 | 0.24353 | 0.26281 |
| V6SH | 0.0070 | 0.3496 | 0.3352 | 0.0040 | 0.1717 | <.0001 | 0.0002 | 0.2209 | 0.1854 |
| O6SH | 0.42819 | 0.08669 | -0.30450 | -0.15439 | -0.18170 | 0.15060 | -0.00516 | 0.25900 | 0.30674 |
| O6SH | 0.0259 | 0.6672 | 0.1225 | 0.4419 | 0.3644 | 0.4534 | 0.9796 | 0.1921 | 0.1196 |
| V8SH | 0.68634 | -0.02391 | 0.09221 | 0.24697 | -0.03545 | 0.25708 | 0.59466 | 0.44764 | 0.42169 |
| V8SH | <.0001 | 0.9058 | 0.6474 | 0.2143 | 0.8607 | 0.1955 | 0.0011 | 0.0192 | 0.0285 |
| O8SH | 0.71289 | -0.06367 | 0.13231 | 0.16997 | -0.18556 | 0.00807 | 0.60408 | 0.56938 | 0.48009 |
| O8SH | <.0001 | 0.7524 | 0.5106 | 0.3967 | 0.3541 | 0.9681 | 0.0008 | 0.0019 | 0.0113 |
| V2BLMT | 0.80548 | 0.07781 | -0.07329 | -0.18835 | -0.49274 | -0.28407 | 0.14594 | 0.73987 | 0.64506 |
| V2BLMT | <.0001 | 0.6997 | 0.7164 | 0.3468 | 0.0090 | 0.1510 | 0.4676 | <.0001 | 0.0003 |
| O2BLMT | 1.00000 | 0.28813 | 0.17423 | 0.11529 | -0.07232 | 0.15587 | 0.32226 | 0.80326 | 0.74381 |
| O2BLMT | | 0.1450 | 0.3847 | 0.5669 | 0.7200 | 0.4375 | 0.1011 | <.0001 | <.0001 |
| V4BLMT | 0.28813 | 1.00000 | 0.61511 | -0.03408 | 0.23123 | 0.06968 | -0.31702 | 0.51744 | 0.56386 |
| V4BLMT | 0.1450 | | 0.0006 | 0.8660 | 0.2459 | 0.7298 | 0.1071 | 0.0057 | 0.0022 |
| O4BLMT | 0.17423 | 0.61511 | 1.00000 | 0.08202 | 0.22237 | 0.09356 | 0.13626 | 0.17094 | 0.13989 |
| O4BLMT | 0.3847 | 0.0006 | | 0.6842 | 0.2649 | 0.6425 | 0.4980 | 0.3939 | 0.4865 |
| V6BLMT | 0.11529 | -0.03408 | 0.08202 | 1.00000 | 0.78723 | 0.84377 | 0.64101 | 0.08626 | 0.11662 |
| V6BLMT | 0.5669 | 0.8660 | 0.6842 | | <.0001 | <.0001 | 0.0003 | 0.6688 | 0.5624 |
| O6BLMT | -0.07232 | 0.23123 | 0.22237 | 0.78723 | 1.00000 | 0.77823 | 0.25091 | -0.06354 | 0.06640 |
| O6BLMT | 0.7200 | 0.2459 | 0.2649 | <.0001 | | <.0001 | 0.2068 | 0.7529 | 0.7421 |
| V6ND | 0.15587 | 0.06968 | 0.09356 | 0.84377 | 0.77823 | 1.00000 | 0.58820 | 0.03613 | 0.10180 |
| V6ND | 0.4375 | 0.7298 | 0.6425 | <.0001 | <.0001 | | 0.0013 | 0.8580 | 0.6134 |

| Variable | V2SH | O2SH | V4SH | O4SH | V6SH | O6SH | V8SH | O8SH | V2BLMT |
|---|---|---|---|---|---|---|---|---|---|
| O6ND | -0.12833 | 0.05510 | -0.16629 | -0.23806 | 0.65370 | -0.00516 | 0.59466 | 0.60408 | 0.14594 |
| O6ND | 0.5235 | 0.7849 | 0.4071 | 0.2318 | 0.0002 | 0.9796 | 0.0011 | 0.0008 | 0.4676 |
| V8ND | 0.67514 | 0.69602 | -0.07609 | 0.07822 | 0.24353 | 0.25900 | 0.44764 | 0.56938 | 0.73987 |
| V8ND | 0.0001 | <.0001 | 0.7060 | 0.6982 | 0.2209 | 0.1921 | 0.0192 | 0.0019 | <.0001 |
| O8ND | 0.65776 | 0.67795 | 0.09404 | 0.26382 | 0.26281 | 0.30674 | 0.42169 | 0.48009 | 0.64506 |
| O8ND | 0.0002 | 0.0001 | 0.6408 | 0.1836 | 0.1854 | 0.1196 | 0.0285 | 0.0113 | 0.0003 |

CLUSTER = 2

| Variable | O2BLMT | V4BLMT | O4BLMT | V6BLMT | O6BLMT | V6ND | O6ND | V8ND | O8ND |
|---|---|---|---|---|---|---|---|---|---|
| O6ND | 0.32226 | -0.31702 | 0.13626 | 0.64101 | 0.25091 | 0.58820 | 1.00000 | 0.10060 | 0.00962 |
| O6ND | 0.1011 | 0.1071 | 0.4980 | 0.0003 | 0.2068 | 0.0013 | | 0.6176 | 0.9620 |
| V8ND | 0.80326 | 0.51744 | 0.17094 | 0.08626 | -0.06354 | 0.03613 | 0.10060 | 1.00000 | 0.95388 |
| V8ND | <.0001 | 0.0057 | 0.3939 | 0.6688 | 0.7529 | 0.8580 | 0.6176 | | <.0001 |
| O8ND | 0.74381 | 0.56386 | 0.13989 | 0.11662 | 0.06640 | 0.10180 | 0.00962 | 0.95388 | 1.00000 |
| O8ND | <.0001 | 0.0022 | 0.4865 | 0.5624 | 0.7421 | 0.6134 | 0.9620 | <.0001 | |

***Cluster 3***

| Variable | V2SH | O2SH | V4SH | O4SH | V6SH | O6SH | V8SH | O8SH | V2BLMT |
|---|---|---|---|---|---|---|---|---|---|
| V2SH | 1.00000 | 0.96698 | 0.52901 | 0.25833 | 0.70383 | 0.53109 | 0.94926 | 0.94557 | 0.90085 |
| V2SH | | <.0001 | 0.0770 | 0.4175 | 0.0106 | 0.0756 | <.0001 | <.0001 | <.0001 |
| O2SH | 0.96698 | 1.00000 | 0.62668 | 0.37290 | 0.57979 | 0.42006 | 0.91795 | 0.95277 | 0.83508 |
| O2SH | <.0001 | | 0.0292 | 0.2325 | 0.0482 | 0.1740 | <.0001 | <.0001 | 0.0007 |
| V4SH | 0.52901 | 0.62668 | 1.00000 | 0.75882 | 0.29508 | 0.00527 | 0.62976 | 0.60218 | 0.49902 |
| V4SH | 0.0770 | 0.0292 | | 0.0042 | 0.3518 | 0.9870 | 0.0282 | 0.0383 | 0.0986 |
| O4SH | 0.25833 | 0.37290 | 0.75882 | 1.00000 | 0.32546 | 0.20855 | 0.34459 | 0.34037 | 0.42636 |
| O4SH | 0.4175 | 0.2325 | 0.0042 | | 0.3019 | 0.5154 | 0.2727 | 0.2790 | 0.1669 |
| V6SH | 0.70383 | 0.57979 | 0.29508 | 0.32546 | 1.00000 | 0.85448 | 0.74388 | 0.66691 | 0.88494 |
| V6SH | 0.0106 | 0.0482 | 0.3518 | 0.3019 | | 0.0004 | 0.0055 | 0.0178 | 0.0001 |
| O6SH | 0.53109 | 0.42006 | 0.00527 | 0.20855 | 0.85448 | 1.00000 | 0.54415 | 0.51418 | 0.74561 |
| O6SH | 0.0756 | 0.1740 | 0.9870 | 0.5154 | 0.0004 | | 0.0674 | 0.0872 | 0.0054 |
| V8SH | 0.94926 | 0.91795 | 0.62976 | 0.34459 | 0.74388 | 0.54415 | 1.00000 | 0.97786 | 0.93492 |
| V8SH | <.0001 | <.0001 | 0.0282 | 0.2727 | 0.0055 | 0.0674 | | <.0001 | <.0001 |
| O8SH | 0.94557 | 0.95277 | 0.60218 | 0.34037 | 0.66691 | 0.51418 | 0.97786 | 1.00000 | 0.89400 |
| O8SH | <.0001 | <.0001 | 0.0383 | 0.2790 | 0.0178 | 0.0872 | <.0001 | | <.0001 |
| V2BLMT | 0.90085 | 0.83508 | 0.49902 | 0.42636 | 0.88494 | 0.74561 | 0.93492 | 0.89400 | 1.00000 |
| V2BLMT | <.0001 | 0.0007 | 0.0986 | 0.1669 | 0.0001 | 0.0054 | <.0001 | <.0001 | |
| O2BLMT | 0.69433 | 0.64083 | 0.45340 | 0.57267 | 0.87423 | 0.80739 | 0.74792 | 0.71861 | 0.87533 |
| O2BLMT | 0.0122 | 0.0247 | 0.1388 | 0.0516 | 0.0002 | 0.0015 | 0.0052 | 0.0085 | 0.0002 |
| V4BLMT | 0.42067 | 0.48213 | 0.36694 | -0.15190 | -0.22773 | -0.26824 | 0.37885 | 0.41507 | 0.10034 |
| V4BLMT | 0.1733 | 0.1124 | 0.2407 | 0.6375 | 0.4766 | 0.3992 | 0.2246 | 0.1797 | 0.7564 |
| O4BLMT | 0.49098 | 0.42116 | 0.29139 | -0.06371 | 0.18582 | 0.21302 | 0.50983 | 0.44352 | 0.40165 |
| O4BLMT | 0.1050 | 0.1727 | 0.3581 | 0.8441 | 0.5631 | 0.5062 | 0.0904 | 0.1487 | 0.1956 |
| V6BLMT | 0.49626 | 0.55280 | 0.32988 | -0.23065 | -0.09423 | -0.25445 | 0.39483 | 0.43374 | 0.12605 |
| V6BLMT | 0.1008 | 0.0623 | 0.2950 | 0.4708 | 0.7708 | 0.4248 | 0.2040 | 0.1589 | 0.6963 |
| O6BLMT | 0.42998 | 0.48543 | 0.30159 | -0.24873 | -0.09842 | -0.28714 | 0.37928 | 0.40725 | 0.11227 |
| O6BLMT | 0.1630 | 0.1096 | 0.3408 | 0.4356 | 0.7609 | 0.3655 | 0.2240 | 0.1889 | 0.7283 |
| V6ND | 0.46979 | 0.52884 | 0.32496 | -0.22256 | -0.12060 | -0.29168 | 0.35612 | 0.39267 | 0.09547 |
| V6ND | 0.1233 | 0.0771 | 0.3027 | 0.4869 | 0.7089 | 0.3576 | 0.2559 | 0.2067 | 0.7679 |

Cluster 3 (Con't.)

| Variable | O2BLMT | V4BLMT | O4BLMT | V6BLMT | O6BLMT | V6ND | O6ND | V8ND | O8ND |
|---|---|---|---|---|---|---|---|---|---|
| V2SH | 0.69433 | 0.42067 | 0.49098 | 0.49626 | 0.42998 | 0.46979 | 0.28103 | 0.43951 | 0.40753 |
| V2SH | 0.0122 | 0.1733 | 0.1050 | 0.1008 | 0.1630 | 0.1233 | 0.3763 | 0.1528 | 0.1885 |
| O2SH | 0.64083 | 0.48213 | 0.42116 | 0.55280 | 0.48543 | 0.52884 | 0.37913 | 0.54815 | 0.50965 |
| O2SH | 0.0247 | 0.1124 | 0.1727 | 0.0623 | 0.1096 | 0.0771 | 0.2242 | 0.0650 | 0.0905 |
| V4SH | 0.45340 | 0.36694 | 0.29139 | 0.32988 | 0.30159 | 0.32496 | 0.23275 | 0.35124 | 0.32454 |
| V4SH | 0.1388 | 0.2407 | 0.3581 | 0.2950 | 0.3408 | 0.3027 | 0.4666 | 0.2629 | 0.3034 |
| O4SH | 0.57267 | -0.15190 | -0.06371 | -0.23065 | -0.24873 | -0.22256 | -0.28042 | -0.14102 | -0.17176 |
| O4SH | 0.0516 | 0.6375 | 0.8441 | 0.4708 | 0.4356 | 0.4869 | 0.3773 | 0.6620 | 0.5935 |
| V6SH | 0.87423 | -0.22773 | 0.18582 | -0.09423 | -0.09842 | -0.12060 | -0.38799 | -0.29385 | -0.32145 |
| V6SH | 0.0002 | 0.4766 | 0.5631 | 0.7708 | 0.7609 | 0.7089 | 0.2127 | 0.3539 | 0.3083 |
| O6SH | 0.80739 | -0.26824 | 0.21302 | -0.25445 | -0.28714 | -0.29168 | -0.48474 | -0.38957 | -0.42079 |
| O6SH | 0.0015 | 0.3992 | 0.5062 | 0.4248 | 0.3655 | 0.3576 | 0.1102 | 0.2106 | 0.1732 |
| V8SH | 0.74792 | 0.37885 | 0.50983 | 0.39483 | 0.37928 | 0.35612 | 0.17316 | 0.35111 | 0.31133 |
| V8SH | 0.0052 | 0.2246 | 0.0904 | 0.2040 | 0.2240 | 0.2559 | 0.5904 | 0.2631 | 0.3246 |
| O8SH | 0.71861 | 0.41507 | 0.44352 | 0.43374 | 0.40725 | 0.39267 | 0.24075 | 0.43240 | 0.38530 |
| O8SH | 0.0085 | 0.1797 | 0.1487 | 0.1589 | 0.1889 | 0.2067 | 0.4510 | 0.1604 | 0.2161 |
| V2BLMT | 0.87533 | 0.10034 | 0.40165 | 0.12605 | 0.11227 | 0.09547 | -0.11166 | 0.06808 | 0.02895 |
| V2BLMT | 0.0002 | 0.7564 | 0.1956 | 0.6963 | 0.7283 | 0.7679 | 0.7297 | 0.8335 | 0.9288 |
| O2BLMT | 1.00000 | -0.11559 | 0.23280 | -0.21324 | -0.24124 | -0.24833 | -0.43250 | -0.24856 | -0.29274 |
| O2BLMT | | 0.7206 | 0.4665 | 0.5058 | 0.4500 | 0.4364 | 0.1603 | 0.4360 | 0.3558 |
| V4BLMT | -0.11559 | 1.00000 | 0.73467 | 0.78402 | 0.74997 | 0.74978 | 0.79301 | 0.86518 | 0.87170 |
| V4BLMT | 0.7206 | | 0.0065 | 0.0025 | 0.0050 | 0.0050 | 0.0021 | 0.0003 | 0.0002 |
| O4BLMT | 0.23280 | 0.73467 | 1.00000 | 0.37875 | 0.34209 | 0.32900 | 0.27868 | 0.39490 | 0.40862 |
| O4BLMT | 0.4665 | 0.0065 | | 0.2247 | 0.2764 | 0.2964 | 0.3804 | 0.2039 | 0.1872 |
| V6BLMT | -0.21324 | 0.78402 | 0.37875 | 1.00000 | 0.95503 | 0.99580 | 0.94400 | 0.90930 | 0.91633 |
| V6BLMT | 0.5058 | 0.0025 | 0.2247 | | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 |
| O6BLMT | -0.24124 | 0.74997 | 0.34209 | 0.95503 | 1.00000 | 0.94533 | 0.90634 | 0.86305 | 0.87032 |
| O6BLMT | 0.4500 | 0.0050 | 0.2764 | <.0001 | | <.0001 | <.0001 | 0.0003 | 0.0002 |
| V6ND | -0.24833 | 0.74978 | 0.32900 | 0.99580 | 0.94533 | 1.00000 | 0.95152 | 0.90323 | 0.91149 |
| V6ND | 0.4364 | 0.0050 | 0.2964 | <.0001 | <.0001 | | <.0001 | <.0001 | <.0001 |

| Variable | V2SH | O2SH | V4SH | O4SH | V6SH | O6SH | V8SH | O8SH | V2BLMT |
|---|---|---|---|---|---|---|---|---|---|
| O6ND | 0.28103 | 0.37913 | 0.23275 | -0.28042 | -0.38799 | -0.48474 | 0.17316 | 0.24075 | -0.11166 |
| O6ND | 0.3763 | 0.2242 | 0.4666 | 0.3773 | 0.2127 | 0.1102 | 0.5904 | 0.4510 | 0.7297 |
| V8ND | 0.43951 | 0.54815 | 0.35124 | -0.14102 | -0.29385 | -0.38957 | 0.35111 | 0.43240 | 0.06808 |
| V8ND | 0.1528 | 0.0650 | 0.2629 | 0.6620 | 0.3539 | 0.2106 | 0.2631 | 0.1604 | 0.8335 |
| O8ND | 0.40753 | 0.50965 | 0.32454 | -0.17176 | -0.32145 | -0.42079 | 0.31133 | 0.38530 | 0.02895 |
| O8ND | 0.1885 | 0.0905 | 0.3034 | 0.5935 | 0.3083 | 0.1732 | 0.3246 | 0.2161 | 0.9288 |

CLUSTER = 3

| Variable | O2BLMT | V4BLMT | O4BLMT | V6BLMT | O6BLMT | V6ND | O6ND | V8ND | O8ND |
|---|---|---|---|---|---|---|---|---|---|
| O6ND | -0.43250 | 0.79301 | 0.27868 | 0.94400 | 0.90634 | 0.95152 | 1.00000 | 0.95422 | 0.96119 |
| O6ND | 0.1603 | 0.0021 | 0.3804 | <.0001 | <.0001 | <.0001 | | <.0001 | <.0001 |
| V8ND | -0.24856 | 0.86518 | 0.39490 | 0.90930 | 0.86305 | 0.90323 | 0.95422 | 1.00000 | 0.99655 |
| V8ND | 0.4360 | 0.0003 | 0.2039 | <.0001 | 0.0003 | <.0001 | <.0001 | | <.0001 |
| O8ND | -0.29274 | 0.87170 | 0.40862 | 0.91633 | 0.87032 | 0.91149 | 0.96119 | 0.99655 | 1.00000 |
| O8ND | 0.3558 | 0.0002 | 0.1872 | <.0001 | 0.0002 | <.0001 | <.0001 | <.0001 | |

***Cluster 4***

| Variable | V2SH | O2SH | V4SH | O4SH | V6SH | O6SH | V8SH | O8SH | V2BLMT |
|---|---|---|---|---|---|---|---|---|---|
| V2SH | 1.00000 | 0.98026 | 0.93647 | 0.80387 | 0.92136 | 0.56493 | 0.66234 | 0.66324 | 0.90281 |
| V2SH | | 0.0001 | 0.0019 | 0.0294 | 0.0032 | 0.1864 | 0.1050 | 0.1044 | 0.0054 |
| O2SH | 0.98026 | 1.00000 | 0.91839 | 0.75354 | 0.93744 | 0.51289 | 0.63323 | 0.63133 | 0.85981 |
| O2SH | 0.0001 | | 0.0035 | 0.0505 | 0.0018 | 0.2391 | 0.1269 | 0.1284 | 0.0131 |
| V4SH | 0.93647 | 0.91839 | 1.00000 | 0.84928 | 0.76111 | 0.38053 | 0.38486 | 0.40144 | 0.71050 |
| V4SH | 0.0019 | 0.0035 | | 0.0156 | 0.0469 | 0.3997 | 0.3939 | 0.3721 | 0.0736 |
| O4SH | 0.80387 | 0.75354 | 0.84928 | 1.00000 | 0.55434 | 0.16352 | 0.47559 | 0.52581 | 0.70513 |
| O4SH | 0.0294 | 0.0505 | 0.0156 | | 0.1966 | 0.7261 | 0.2807 | 0.2255 | 0.0768 |
| V6SH | 0.92136 | 0.93744 | 0.76111 | 0.55434 | 1.00000 | 0.61168 | 0.77976 | 0.76012 | 0.90215 |
| V6SH | 0.0032 | 0.0018 | 0.0469 | 0.1966 | | 0.1444 | 0.0387 | 0.0473 | 0.0055 |
| O6SH | 0.56493 | 0.51289 | 0.38053 | 0.16352 | 0.61168 | 1.00000 | 0.46242 | 0.39342 | 0.63939 |
| O6SH | 0.1864 | 0.2391 | 0.3997 | 0.7261 | 0.1444 | | 0.2961 | 0.3826 | 0.1221 |
| V8SH | 0.66234 | 0.63323 | 0.38486 | 0.47559 | 0.77976 | 0.46242 | 1.00000 | 0.99489 | 0.90244 |
| V8SH | 0.1050 | 0.1269 | 0.3939 | 0.2807 | 0.0387 | 0.2961 | | <.0001 | 0.0054 |
| O8SH | 0.66324 | 0.63133 | 0.40144 | 0.52581 | 0.76012 | 0.39342 | 0.99489 | 1.00000 | 0.89472 |
| O8SH | 0.1044 | 0.1284 | 0.3721 | 0.2255 | 0.0473 | 0.3826 | <.0001 | | 0.0065 |
| V2BLMT | 0.90281 | 0.85981 | 0.71050 | 0.70513 | 0.90215 | 0.63939 | 0.90244 | 0.89472 | 1.00000 |
| V2BLMT | 0.0054 | 0.0131 | 0.0736 | 0.0768 | 0.0055 | 0.1221 | 0.0054 | 0.0065 | |
| O2BLMT | 0.92310 | 0.90841 | 0.74264 | 0.68693 | 0.95021 | 0.57774 | 0.89089 | 0.88281 | 0.98389 |
| O2BLMT | 0.0030 | 0.0046 | 0.0559 | 0.0882 | 0.0010 | 0.1743 | 0.0071 | 0.0085 | <.0001 |
| V4BLMT | -0.20828 | -0.10507 | -0.22199 | -0.52118 | 0.01713 | -0.13744 | -0.12907 | -0.18094 | -0.23780 |
| V4BLMT | 0.6540 | 0.8226 | 0.6323 | 0.2303 | 0.9709 | 0.7689 | 0.7827 | 0.6978 | 0.6076 |
| O4BLMT | -0.43175 | -0.43861 | -0.55290 | -0.46921 | -0.36154 | 0.30113 | -0.10549 | -0.17094 | -0.19472 |
| O4BLMT | 0.3334 | 0.3249 | 0.1980 | 0.2882 | 0.4255 | 0.5116 | 0.8219 | 0.7140 | 0.6757 |
| V6BLMT | 0.42578 | 0.56651 | 0.37057 | 0.11300 | 0.52343 | 0.38230 | 0.18298 | 0.13220 | 0.31796 |
| V6BLMT | 0.3409 | 0.1849 | 0.4132 | 0.8094 | 0.2279 | 0.3974 | 0.6945 | 0.7775 | 0.4871 |
| O6BLMT | 0.71393 | 0.83084 | 0.67454 | 0.48954 | 0.75106 | 0.26498 | 0.42736 | 0.41371 | 0.57663 |
| O6BLMT | 0.0716 | 0.0206 | 0.0965 | 0.2648 | 0.0517 | 0.5658 | 0.3389 | 0.3562 | 0.1754 |
| V6ND | 0.08105 | 0.20779 | 0.25616 | 0.05185 | -0.00780 | -0.01293 | -0.44469 | -0.46728 | -0.20181 |
| V6ND | 0.8629 | 0.6548 | 0.5793 | 0.9121 | 0.9868 | 0.9781 | 0.3174 | 0.2904 | 0.6643 |

Cluster 4 (Con't.)

| Variable | O2BLMT | V4BLMT | O4BLMT | V6BLMT | O6BLMT | V6ND | O6ND | V8ND | O8ND |
|---|---|---|---|---|---|---|---|---|---|
| V2SH | 0.92310 | -0.20828 | -0.43175 | 0.42578 | 0.71393 | 0.08105 | 0.13425 | 0.80130 | 0.65993 |
| V2SH | 0.0030 | 0.6540 | 0.3334 | 0.3409 | 0.0716 | 0.8629 | 0.7741 | 0.0303 | 0.1067 |
| O2SH | 0.90841 | -0.10507 | -0.43861 | 0.56651 | 0.83084 | 0.20779 | 0.26618 | 0.85625 | 0.74003 |
| O2SH | 0.0046 | 0.8226 | 0.3249 | 0.1849 | 0.0206 | 0.6548 | 0.5639 | 0.0139 | 0.0572 |
| V4SH | 0.74264 | -0.22199 | -0.55290 | 0.37057 | 0.67454 | 0.25616 | 0.30247 | 0.74133 | 0.62220 |
| V4SH | 0.0559 | 0.6323 | 0.1980 | 0.4132 | 0.0965 | 0.5793 | 0.5097 | 0.0565 | 0.1357 |
| O4SH | 0.68693 | -0.52118 | -0.46921 | 0.11300 | 0.48954 | 0.05185 | -0.02549 | 0.44479 | 0.27836 |
| O4SH | 0.0882 | 0.2303 | 0.2882 | 0.8094 | 0.2648 | 0.9121 | 0.9567 | 0.3173 | 0.5455 |
| V6SH | 0.95021 | 0.01713 | -0.36154 | 0.52343 | 0.75106 | -0.00780 | 0.10397 | 0.89412 | 0.79195 |
| V6SH | 0.0010 | 0.9709 | 0.4255 | 0.2279 | 0.0517 | 0.9868 | 0.8245 | 0.0066 | 0.0338 |
| O6SH | 0.57774 | -0.13744 | 0.30113 | 0.38230 | 0.26498 | -0.01293 | 0.01484 | 0.28873 | 0.17038 |
| O6SH | 0.1743 | 0.7689 | 0.5116 | 0.3974 | 0.5658 | 0.9781 | 0.9748 | 0.5300 | 0.7149 |
| V8SH | 0.89089 | -0.12907 | -0.10549 | 0.18298 | 0.42736 | -0.44469 | -0.41678 | 0.54338 | 0.41847 |
| V8SH | 0.0071 | 0.7827 | 0.8219 | 0.6945 | 0.3389 | 0.3174 | 0.3523 | 0.2074 | 0.3501 |
| O8SH | 0.88281 | -0.18094 | -0.17094 | 0.13220 | 0.41371 | -0.46728 | -0.44840 | 0.53935 | 0.40777 |
| O8SH | 0.0085 | 0.6978 | 0.7140 | 0.7775 | 0.3562 | 0.2904 | 0.3129 | 0.2115 | 0.3638 |
| V2BLMT | 0.98389 | -0.23780 | -0.19472 | 0.31796 | 0.57663 | -0.20181 | -0.17942 | 0.65613 | 0.49717 |
| V2BLMT | <.0001 | 0.6076 | 0.6757 | 0.4871 | 0.1754 | 0.6643 | 0.7003 | 0.1095 | 0.2563 |
| O2BLMT | 1.00000 | -0.10939 | -0.26613 | 0.41653 | 0.68798 | -0.11576 | -0.06388 | 0.76507 | 0.63350 |
| O2BLMT | | 0.8154 | 0.5640 | 0.3526 | 0.0875 | 0.8048 | 0.8918 | 0.0451 | 0.1266 |
| V4BLMT | -0.10939 | 1.00000 | 0.20750 | 0.42993 | 0.26900 | 0.32796 | 0.50821 | 0.23175 | 0.44449 |
| V4BLMT | 0.8154 | | 0.6553 | 0.3357 | 0.5597 | 0.4727 | 0.2442 | 0.6170 | 0.3177 |
| O4BLMT | -0.26613 | 0.20750 | 1.00000 | 0.17051 | -0.23263 | 0.13462 | 0.01774 | -0.62659 | -0.57739 |
| O4BLMT | 0.5640 | 0.6553 | | 0.7147 | 0.6157 | 0.7735 | 0.9699 | 0.1321 | 0.1747 |
| V6BLMT | 0.41653 | 0.42993 | 0.17051 | 1.00000 | 0.87401 | 0.74996 | 0.76892 | 0.48939 | 0.52969 |
| V6BLMT | 0.3526 | 0.3357 | 0.7147 | | 0.0101 | 0.0522 | 0.0433 | 0.2650 | 0.2214 |
| O6BLMT | 0.68798 | 0.26900 | -0.23263 | 0.87401 | 1.00000 | 0.57239 | 0.61326 | 0.76350 | 0.75290 |
| O6BLMT | 0.0875 | 0.5597 | 0.6157 | 0.0101 | | 0.1793 | 0.1431 | 0.0458 | 0.0508 |
| V6ND | -0.11576 | 0.32796 | 0.13462 | 0.74996 | 0.57239 | 1.00000 | 0.95655 | 0.09905 | 0.18350 |
| V6ND | 0.8048 | 0.4727 | 0.7735 | 0.0522 | 0.1793 | | 0.0007 | 0.8327 | 0.6937 |

| Variable | V2SH | O2SH | V4SH | O4SH | V6SH | O6SH | V8SH | O8SH | V2BLMT |
|---|---|---|---|---|---|---|---|---|---|
| O6ND | 0.13425 | 0.26618 | 0.30247 | -0.02549 | 0.10397 | 0.01484 | -0.41678 | -0.44840 | -0.17942 |
| O6ND | 0.7741 | 0.5639 | 0.5097 | 0.9567 | 0.8245 | 0.9748 | 0.3523 | 0.3129 | 0.7003 |
| V8ND | 0.80130 | 0.85625 | 0.74133 | 0.44479 | 0.89412 | 0.28873 | 0.54338 | 0.53935 | 0.65613 |
| V8ND | 0.0303 | 0.0139 | 0.0565 | 0.3173 | 0.0066 | 0.5300 | 0.2074 | 0.2115 | 0.1095 |
| O8ND | 0.65993 | 0.74003 | 0.62220 | 0.27836 | 0.79195 | 0.17038 | 0.41847 | 0.40777 | 0.49717 |
| O8ND | 0.1067 | 0.0572 | 0.1357 | 0.5455 | 0.0338 | 0.7149 | 0.3501 | 0.3638 | 0.2563 |

CLUSTER = 4

| Variable | O2BLMT | V4BLMT | O4BLMT | V6BLMT | O6BLMT | V6ND | O6ND | V8ND | O8ND |
|---|---|---|---|---|---|---|---|---|---|
| O6ND | -0.06388 | 0.50821 | 0.01774 | 0.76892 | 0.61326 | 0.95655 | 1.00000 | 0.28232 | 0.39626 |
| O6ND | 0.8918 | 0.2442 | 0.9699 | 0.0433 | 0.1431 | 0.0007 | | 0.5396 | 0.3788 |
| V8ND | 0.76507 | 0.23175 | -0.62659 | 0.48939 | 0.76350 | 0.09905 | 0.28232 | 1.00000 | 0.97081 |
| V8ND | 0.0451 | 0.6170 | 0.1321 | 0.2650 | 0.0458 | 0.8327 | 0.5396 | | 0.0003 |
| O8ND | 0.63350 | 0.44449 | -0.57739 | 0.52969 | 0.75290 | 0.18350 | 0.39626 | 0.97081 | 1.00000 |
| O8ND | 0.1266 | 0.3177 | 0.1747 | 0.2214 | 0.0508 | 0.6937 | 0.3788 | 0.0003 | |

***Cluster 5***

| Variable | V2SH | O2SH | V4SH | O4SH | V6SH | O6SH | V8SH | O8SH | V2BLMT |
|---|---|---|---|---|---|---|---|---|---|
| V2SH | 1.00000 | 0.65067 | 0.53905 | 0.39102 | 0.68311 | 0.15255 | 0.90721 | 0.56768 | 0.90725 |
| V2SH | | 0.0008 | 0.0080 | 0.0650 | 0.0003 | 0.4871 | <.0001 | 0.0047 | <.0001 |
| O2SH | 0.65067 | 1.00000 | 0.17030 | 0.07768 | 0.34010 | 0.29381 | 0.32481 | 0.14811 | 0.32796 |
| O2SH | 0.0008 | | 0.4372 | 0.7246 | 0.1123 | 0.1736 | 0.1305 | 0.5000 | 0.1266 |
| V4SH | 0.53905 | 0.17030 | 1.00000 | 0.88220 | 0.52549 | -0.01061 | 0.66117 | 0.51814 | 0.55734 |
| V4SH | 0.0080 | 0.4372 | | <.0001 | 0.0100 | 0.9617 | 0.0006 | 0.0113 | 0.0057 |
| O4SH | 0.39102 | 0.07768 | 0.88220 | 1.00000 | 0.38042 | -0.10548 | 0.49581 | 0.42852 | 0.42281 |
| O4SH | 0.0650 | 0.7246 | <.0001 | | 0.0733 | 0.6320 | 0.0161 | 0.0413 | 0.0444 |
| V6SH | 0.68311 | 0.34010 | 0.52549 | 0.38042 | 1.00000 | 0.60888 | 0.73229 | 0.43599 | 0.69694 |
| V6SH | 0.0003 | 0.1123 | 0.0100 | 0.0733 | | 0.0020 | <.0001 | 0.0376 | 0.0002 |
| O6SH | 0.15255 | 0.29381 | -0.01061 | -0.10548 | 0.60888 | 1.00000 | 0.02791 | -0.12572 | 0.03197 |
| O6SH | 0.4871 | 0.1736 | 0.9617 | 0.6320 | 0.0020 | | 0.8994 | 0.5676 | 0.8849 |
| V8SH | 0.90721 | 0.32481 | 0.66117 | 0.49581 | 0.73229 | 0.02791 | 1.00000 | 0.65589 | 0.94512 |
| V8SH | <.0001 | 0.1305 | 0.0006 | 0.0161 | <.0001 | 0.8994 | | 0.0007 | <.0001 |
| O8SH | 0.56768 | 0.14811 | 0.51814 | 0.42852 | 0.43599 | -0.12572 | 0.65589 | 1.00000 | 0.65733 |
| O8SH | 0.0047 | 0.5000 | 0.0113 | 0.0413 | 0.0376 | 0.5676 | 0.0007 | | 0.0007 |
| V2BLMT | 0.90725 | 0.32796 | 0.55734 | 0.42281 | 0.69694 | 0.03197 | 0.94512 | 0.65733 | 1.00000 |
| V2BLMT | <.0001 | 0.1266 | 0.0057 | 0.0444 | 0.0002 | 0.8849 | <.0001 | 0.0007 | |
| O2BLMT | 0.69967 | 0.34279 | 0.26029 | 0.15506 | 0.54870 | 0.16206 | 0.61932 | 0.43385 | 0.81129 |
| O2BLMT | 0.0002 | 0.1093 | 0.2303 | 0.4799 | 0.0067 | 0.4601 | 0.0016 | 0.0386 | <.0001 |
| V4BLMT | 0.78291 | 0.50023 | 0.08886 | 0.07095 | 0.48632 | 0.23735 | 0.61278 | 0.50638 | 0.72802 |
| V4BLMT | <.0001 | 0.0151 | 0.6868 | 0.7477 | 0.0186 | 0.2755 | 0.0019 | 0.0137 | <.0001 |
| O4BLMT | 0.84998 | 0.53332 | 0.21226 | 0.17691 | 0.56037 | 0.21660 | 0.70297 | 0.58906 | 0.79615 |
| O4BLMT | <.0001 | 0.0088 | 0.3309 | 0.4194 | 0.0054 | 0.3209 | 0.0002 | 0.0031 | <.0001 |
| V6BLMT | 0.90604 | 0.36961 | 0.43371 | 0.34746 | 0.71490 | 0.11754 | 0.91845 | 0.65229 | 0.94313 |
| V6BLMT | <.0001 | 0.0826 | 0.0387 | 0.1043 | 0.0001 | 0.5932 | <.0001 | 0.0007 | <.0001 |
| O6BLMT | 0.77558 | 0.30196 | 0.36823 | 0.33900 | 0.74945 | 0.26533 | 0.78157 | 0.55937 | 0.86340 |
| O6BLMT | <.0001 | 0.1614 | 0.0838 | 0.1136 | <.0001 | 0.2211 | <.0001 | 0.0055 | <.0001 |
| V6ND | 0.90967 | 0.41070 | 0.39216 | 0.30110 | 0.72250 | 0.17877 | 0.89340 | 0.62705 | 0.92733 |
| V6ND | <.0001 | 0.0516 | 0.0642 | 0.1627 | <.0001 | 0.4144 | <.0001 | 0.0014 | <.0001 |

Cluster 5 (Con't.)

| Variable | O2BLMT | V4BLMT | O4BLMT | V6BLMT | O6BLMT | V6ND | O6ND | V8ND | O8ND |
|---|---|---|---|---|---|---|---|---|---|
| V2SH | 0.69967 | 0.78291 | 0.84998 | 0.90604 | 0.77558 | 0.90967 | 0.62658 | 0.75779 | 0.58745 |
| V2SH | 0.0002 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | 0.0014 | <.0001 | 0.0032 |
| O2SH | 0.34279 | 0.50023 | 0.53332 | 0.36961 | 0.30196 | 0.41070 | 0.24713 | 0.42610 | 0.23155 |
| O2SH | 0.1093 | 0.0151 | 0.0088 | 0.0826 | 0.1614 | 0.0516 | 0.2556 | 0.0426 | 0.2878 |
| V4SH | 0.26029 | 0.08886 | 0.21226 | 0.43371 | 0.36823 | 0.39216 | 0.11608 | 0.07613 | 0.00565 |
| V4SH | 0.2303 | 0.6868 | 0.3309 | 0.0387 | 0.0838 | 0.0642 | 0.5979 | 0.7299 | 0.9796 |
| O4SH | 0.15506 | 0.07095 | 0.17691 | 0.34746 | 0.33900 | 0.30110 | 0.09919 | 0.06939 | 0.00618 |
| O4SH | 0.4799 | 0.7477 | 0.4194 | 0.1043 | 0.1136 | 0.1627 | 0.6525 | 0.7531 | 0.9777 |
| V6SH | 0.54870 | 0.48632 | 0.56037 | 0.71490 | 0.74945 | 0.72250 | 0.65845 | 0.55671 | 0.57965 |
| V6SH | 0.0067 | 0.0186 | 0.0054 | 0.0001 | <.0001 | <.0001 | 0.0006 | 0.0058 | 0.0037 |
| O6SH | 0.16206 | 0.23735 | 0.21660 | 0.11754 | 0.26533 | 0.17877 | 0.43240 | 0.28817 | 0.39268 |
| O6SH | 0.4601 | 0.2755 | 0.3209 | 0.5932 | 0.2211 | 0.4144 | 0.0393 | 0.1824 | 0.0638 |
| V8SH | 0.61932 | 0.61278 | 0.70297 | 0.91845 | 0.78157 | 0.89340 | 0.60103 | 0.63369 | 0.53554 |
| V8SH | 0.0016 | 0.0019 | 0.0002 | <.0001 | <.0001 | <.0001 | 0.0024 | 0.0012 | 0.0084 |
| O8SH | 0.43385 | 0.50638 | 0.58906 | 0.65229 | 0.55937 | 0.62705 | 0.49092 | 0.51412 | 0.47960 |
| O8SH | 0.0386 | 0.0137 | 0.0031 | 0.0007 | 0.0055 | 0.0014 | 0.0174 | 0.0121 | 0.0206 |
| V2BLMT | 0.81129 | 0.72802 | 0.79615 | 0.94313 | 0.86340 | 0.92733 | 0.63187 | 0.73521 | 0.61084 |
| V2BLMT | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | 0.0012 | <.0001 | 0.0020 |
| O2BLMT | 1.00000 | 0.71428 | 0.73457 | 0.73607 | 0.77945 | 0.75445 | 0.52086 | 0.68885 | 0.53714 |
| O2BLMT | | 0.0001 | <.0001 | <.0001 | <.0001 | <.0001 | 0.0108 | 0.0003 | 0.0082 |
| V4BLMT | 0.71428 | 1.00000 | 0.98167 | 0.83644 | 0.76351 | 0.87109 | 0.76366 | 0.97793 | 0.81734 |
| V4BLMT | 0.0001 | | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 |
| O4BLMT | 0.73457 | 0.98167 | 1.00000 | 0.88160 | 0.79654 | 0.90690 | 0.74236 | 0.95355 | 0.77555 |
| O4BLMT | <.0001 | <.0001 | | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 |
| V6BLMT | 0.73607 | 0.83644 | 0.88160 | 1.00000 | 0.89393 | 0.99517 | 0.78925 | 0.86977 | 0.77337 |
| V6BLMT | <.0001 | <.0001 | <.0001 | | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 |
| O6BLMT | 0.77945 | 0.76351 | 0.79654 | 0.89393 | 1.00000 | 0.89341 | 0.79511 | 0.81367 | 0.75159 |
| O6BLMT | <.0001 | <.0001 | <.0001 | <.0001 | | <.0001 | <.0001 | <.0001 | <.0001 |
| V6ND | 0.75445 | 0.87109 | 0.90690 | 0.99517 | 0.89341 | 1.00000 | 0.81332 | 0.90208 | 0.80105 |
| V6ND | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | | <.0001 | <.0001 | <.0001 |

| Variable | V2SH | O2SH | V4SH | O4SH | V6SH | O6SH | V8SH | O8SH | V2BLMT |
|---|---|---|---|---|---|---|---|---|---|
| O6ND | 0.62658 | 0.24713 | 0.11608 | 0.09919 | 0.65845 | 0.43240 | 0.60103 | 0.49092 | 0.63187 |
| O6ND | 0.0014 | 0.2556 | 0.5979 | 0.6525 | 0.0006 | 0.0393 | 0.0024 | 0.0174 | 0.0012 |
| V8ND | 0.75779 | 0.42610 | 0.07613 | 0.06939 | 0.55671 | 0.28817 | 0.63369 | 0.51412 | 0.73521 |
| V8ND | <.0001 | 0.0426 | 0.7299 | 0.7531 | 0.0058 | 0.1824 | 0.0012 | 0.0121 | <.0001 |
| O8ND | 0.58745 | 0.23155 | 0.00565 | 0.00618 | 0.57965 | 0.39268 | 0.53554 | 0.47960 | 0.61084 |
| O8ND | 0.0032 | 0.2878 | 0.9796 | 0.9777 | 0.0037 | 0.0638 | 0.0084 | 0.0206 | 0.0020 |

CLUSTER = 5

| Variable | O2BLMT | V4BLMT | O4BLMT | V6BLMT | O6BLMT | V6ND | O6ND | V8ND | O8ND |
|---|---|---|---|---|---|---|---|---|---|
| O6ND | 0.52086 | 0.76366 | 0.74236 | 0.78925 | 0.79511 | 0.81332 | 1.00000 | 0.85939 | 0.96483 |
| O6ND | 0.0108 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | | <.0001 | <.0001 |
| V8ND | 0.68885 | 0.97793 | 0.95355 | 0.86977 | 0.81367 | 0.90208 | 0.85939 | 1.00000 | 0.90475 |
| V8ND | 0.0003 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | | <.0001 |
| O8ND | 0.53714 | 0.81734 | 0.77555 | 0.77337 | 0.75159 | 0.80105 | 0.96483 | 0.90475 | 1.00000 |
| O8ND | 0.0082 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | |

***Cluster 6***

| Variable | V2SH | O2SH | V4SH | O4SH | V6SH | O6SH | V8SH | O8SH | V2BLMT |
|---|---|---|---|---|---|---|---|---|---|
| V2SH | 1.00000 | 0.60524 | 0.71466 | 0.70014 | 0.14634 | 0.65308 | 0.52153 | 0.41082 | 0.74690 |
| V2SH | | 0.2030 | 0.1105 | 0.1214 | 0.7821 | 0.1597 | 0.2886 | 0.4184 | 0.0880 |
| O2SH | 0.60524 | 1.00000 | 0.47809 | 0.72785 | 0.82871 | 0.86918 | 0.41022 | 0.67406 | 0.19126 |
| O2SH | 0.2030 | | 0.3375 | 0.1010 | 0.0415 | 0.0246 | 0.4192 | 0.1420 | 0.7166 |
| V4SH | 0.71466 | 0.47809 | 1.00000 | 0.91833 | 0.36149 | 0.77767 | 0.92595 | 0.83203 | 0.84029 |
| V4SH | 0.1105 | 0.3375 | | 0.0097 | 0.4814 | 0.0686 | 0.0080 | 0.0400 | 0.0362 |
| O4SH | 0.70014 | 0.72785 | 0.91833 | 1.00000 | 0.64585 | 0.93008 | 0.84011 | 0.90214 | 0.67575 |
| O4SH | 0.1214 | 0.1010 | 0.0097 | | 0.1659 | 0.0072 | 0.0363 | 0.0139 | 0.1407 |
| V6SH | 0.14634 | 0.82871 | 0.36149 | 0.64585 | 1.00000 | 0.74635 | 0.36774 | 0.78358 | -0.09553 |
| V6SH | 0.7821 | 0.0415 | 0.4814 | 0.1659 | | 0.0883 | 0.4733 | 0.0652 | 0.8571 |
| O6SH | 0.65308 | 0.86918 | 0.77767 | 0.93008 | 0.74635 | 1.00000 | 0.78462 | 0.84654 | 0.56080 |
| O6SH | 0.1597 | 0.0246 | 0.0686 | 0.0072 | 0.0883 | | 0.0646 | 0.0335 | 0.2470 |
| V8SH | 0.52153 | 0.41022 | 0.92595 | 0.84011 | 0.36774 | 0.78462 | 1.00000 | 0.81320 | 0.81414 |
| V8SH | 0.2886 | 0.4192 | 0.0080 | 0.0363 | 0.4733 | 0.0646 | | 0.0491 | 0.0486 |
| O8SH | 0.41082 | 0.67406 | 0.83203 | 0.90214 | 0.78358 | 0.84654 | 0.81320 | 1.00000 | 0.42988 |
| O8SH | 0.4184 | 0.1420 | 0.0400 | 0.0139 | 0.0652 | 0.0335 | 0.0491 | | 0.3949 |
| V2BLMT | 0.74690 | 0.19126 | 0.84029 | 0.67575 | -0.09553 | 0.56080 | 0.81414 | 0.42988 | 1.00000 |
| V2BLMT | 0.0880 | 0.7166 | 0.0362 | 0.1407 | 0.8571 | 0.2470 | 0.0486 | 0.3949 | |
| O2BLMT | 0.47531 | 0.95921 | 0.50941 | 0.74769 | 0.87855 | 0.91893 | 0.54101 | 0.75223 | 0.21725 |
| O2BLMT | 0.3407 | 0.0025 | 0.3020 | 0.0875 | 0.0212 | 0.0096 | 0.2677 | 0.0845 | 0.6793 |
| V4BLMT | 0.74826 | 0.44873 | 0.14698 | 0.22734 | -0.00141 | 0.19614 | -0.16017 | -0.06023 | 0.19137 |
| V4BLMT | 0.0871 | 0.3721 | 0.7811 | 0.6649 | 0.9979 | 0.7096 | 0.7618 | 0.9098 | 0.7165 |
| O4BLMT | 0.33438 | 0.90839 | 0.46040 | 0.68845 | 0.91066 | 0.87043 | 0.53197 | 0.77150 | 0.12237 |
| O4BLMT | 0.5171 | 0.0122 | 0.3582 | 0.1305 | 0.0116 | 0.0241 | 0.2773 | 0.0724 | 0.8174 |
| V6BLMT | 0.06008 | 0.78882 | 0.11710 | 0.42986 | 0.86084 | 0.69018 | 0.25861 | 0.51178 | -0.14207 |
| V6BLMT | 0.9100 | 0.0622 | 0.8252 | 0.3949 | 0.0277 | 0.1291 | 0.6207 | 0.2994 | 0.7883 |
| O6BLMT | 0.53406 | 0.78614 | 0.75986 | 0.91097 | 0.73992 | 0.98116 | 0.82097 | 0.84866 | 0.55827 |
| O6BLMT | 0.2751 | 0.0637 | 0.0796 | 0.0115 | 0.0927 | 0.0005 | 0.0452 | 0.0326 | 0.2496 |
| V6ND | -0.20799 | 0.63391 | -0.16325 | 0.14286 | 0.83835 | 0.34649 | -0.11496 | 0.37366 | -0.57519 |
| V6ND | 0.6925 | 0.1765 | 0.7573 | 0.7872 | 0.0371 | 0.5011 | 0.8283 | 0.4656 | 0.2324 |

Cluster 6 (Con't.)

| Variable | O2BLMT | V4BLMT | O4BLMT | V6BLMT | O6BLMT | V6ND | O6ND | V8ND | O8ND |
|---|---|---|---|---|---|---|---|---|---|
| V2SH | 0.47531 | 0.74826 | 0.33438 | 0.06008 | 0.53406 | -0.20799 | 0.05175 | 0.88796 | 0.95314 |
| V2SH | 0.3407 | 0.0871 | 0.5171 | 0.9100 | 0.2751 | 0.6925 | 0.9224 | 0.0181 | 0.0032 |
| O2SH | 0.95921 | 0.44873 | 0.90839 | 0.78882 | 0.78614 | 0.63391 | 0.72283 | 0.54031 | 0.48640 |
| O2SH | 0.0025 | 0.3721 | 0.0122 | 0.0622 | 0.0637 | 0.1765 | 0.1046 | 0.2684 | 0.3279 |
| V4SH | 0.50941 | 0.14698 | 0.46040 | 0.11710 | 0.75986 | -0.16325 | 0.13605 | 0.74635 | 0.79643 |
| V4SH | 0.3020 | 0.7811 | 0.3582 | 0.8252 | 0.0796 | 0.7573 | 0.7972 | 0.0884 | 0.0579 |
| O4SH | 0.74769 | 0.22734 | 0.68845 | 0.42986 | 0.91097 | 0.14286 | 0.46083 | 0.71250 | 0.68819 |
| O4SH | 0.0875 | 0.6649 | 0.1305 | 0.3949 | 0.0115 | 0.7872 | 0.3577 | 0.1121 | 0.1307 |
| V6SH | 0.87855 | -0.00141 | 0.91066 | 0.86084 | 0.73992 | 0.83835 | 0.91774 | 0.12347 | 0.04844 |
| V6SH | 0.0212 | 0.9979 | 0.0116 | 0.0277 | 0.0927 | 0.0371 | 0.0099 | 0.8157 | 0.9274 |
| O6SH | 0.91893 | 0.19614 | 0.87043 | 0.69018 | 0.98116 | 0.34649 | 0.51034 | 0.74305 | 0.63815 |
| O6SH | 0.0096 | 0.7096 | 0.0241 | 0.1291 | 0.0005 | 0.5011 | 0.3009 | 0.0906 | 0.1727 |
| V8SH | 0.54101 | -0.16017 | 0.53197 | 0.25861 | 0.82097 | -0.11496 | 0.03928 | 0.72127 | 0.67530 |
| V8SH | 0.2677 | 0.7618 | 0.2773 | 0.6207 | 0.0452 | 0.8283 | 0.9411 | 0.1057 | 0.1410 |
| O8SH | 0.75223 | -0.06023 | 0.77150 | 0.51178 | 0.84866 | 0.37366 | 0.58239 | 0.43130 | 0.44476 |
| O8SH | 0.0845 | 0.9098 | 0.0724 | 0.2994 | 0.0326 | 0.4656 | 0.2252 | 0.3932 | 0.3768 |
| V2BLMT | 0.21725 | 0.19137 | 0.12237 | -0.14207 | 0.55827 | -0.57519 | -0.33526 | 0.88120 | 0.86515 |
| V2BLMT | 0.6793 | 0.7165 | 0.8174 | 0.7883 | 0.2496 | 0.2324 | 0.5160 | 0.0203 | 0.0261 |
| O2BLMT | 1.00000 | 0.19001 | 0.98297 | 0.88500 | 0.88226 | 0.66341 | 0.68879 | 0.52926 | 0.41685 |
| O2BLMT | | 0.7184 | 0.0004 | 0.0191 | 0.0200 | 0.1509 | 0.1302 | 0.2802 | 0.4109 |
| V4BLMT | 0.19001 | 1.00000 | 0.03919 | -0.07994 | 0.02731 | -0.05932 | 0.16518 | 0.42526 | 0.55170 |
| V4BLMT | 0.7184 | | 0.9412 | 0.8803 | 0.9590 | 0.9111 | 0.7545 | 0.4006 | 0.2564 |
| O4BLMT | 0.98297 | 0.03919 | 1.00000 | 0.91240 | 0.85008 | 0.73678 | 0.70567 | 0.41240 | 0.30344 |
| O4BLMT | 0.0004 | 0.9412 | | 0.0112 | 0.0320 | 0.0948 | 0.1172 | 0.4165 | 0.5588 |
| V6BLMT | 0.88500 | -0.07994 | 0.91240 | 1.00000 | 0.70623 | 0.83065 | 0.70927 | 0.19697 | -0.01129 |
| V6BLMT | 0.0191 | 0.8803 | 0.0112 | | 0.1168 | 0.0406 | 0.1145 | 0.7084 | 0.9831 |
| O6BLMT | 0.88226 | 0.02731 | 0.85008 | 0.70623 | 1.00000 | 0.32619 | 0.47792 | 0.69331 | 0.54418 |
| O6BLMT | 0.0200 | 0.9590 | 0.0320 | 0.1168 | | 0.5281 | 0.3377 | 0.1267 | 0.2643 |
| V6ND | 0.66341 | -0.05932 | 0.73678 | 0.83065 | 0.32619 | 1.00000 | 0.85314 | -0.26644 | -0.32924 |
| V6ND | 0.1509 | 0.9111 | 0.0948 | 0.0406 | 0.5281 | | 0.0308 | 0.6098 | 0.5240 |

| Variable | V2SH | O2SH | V4SH | O4SH | V6SH | O6SH | V8SH | O8SH | V2BLMT |
|---|---|---|---|---|---|---|---|---|---|
| O6ND | 0.05175 | 0.72283 | 0.13605 | 0.46083 | 0.91774 | 0.51034 | 0.03928 | 0.58239 | -0.33526 |
| O6ND | 0.9224 | 0.1046 | 0.7972 | 0.3577 | 0.0099 | 0.3009 | 0.9411 | 0.2252 | 0.5160 |
| V8ND | 0.88796 | 0.54031 | 0.74635 | 0.71250 | 0.12347 | 0.74305 | 0.72127 | 0.43130 | 0.88120 |
| V8ND | 0.0181 | 0.2684 | 0.0884 | 0.1121 | 0.8157 | 0.0906 | 0.1057 | 0.3932 | 0.0203 |
| O8ND | 0.95314 | 0.48640 | 0.79643 | 0.68819 | 0.04844 | 0.63815 | 0.67530 | 0.44476 | 0.86515 |
| O8ND | 0.0032 | 0.3279 | 0.0579 | 0.1307 | 0.9274 | 0.1727 | 0.1410 | 0.3768 | 0.0261 |

CLUSTER = 6

| Variable | O2BLMT | V4BLMT | O4BLMT | V6BLMT | O6BLMT | V6ND | O6ND | V8ND | O8ND |
|---|---|---|---|---|---|---|---|---|---|
| O6ND | 0.68879 | 0.16518 | 0.70567 | 0.70927 | 0.47792 | 0.85314 | 1.00000 | -0.11947 | -0.13766 |
| O6ND | 0.1302 | 0.7545 | 0.1172 | 0.1145 | 0.3377 | 0.0308 | | 0.8216 | 0.7948 |
| V8ND | 0.52926 | 0.42526 | 0.41240 | 0.19697 | 0.69331 | -0.26644 | -0.11947 | 1.00000 | 0.93347 |
| V8ND | 0.2802 | 0.4006 | 0.4165 | 0.7084 | 0.1267 | 0.6098 | 0.8216 | | 0.0065 |
| O8ND | 0.41685 | 0.55170 | 0.30344 | -0.01129 | 0.54418 | -0.32924 | -0.13766 | 0.93347 | 1.00000 |
| O8ND | 0.4109 | 0.2564 | 0.5588 | 0.9831 | 0.2643 | 0.5240 | 0.7948 | 0.0065 | |

***Cluster 7***

| Variable | V2SH | O2SH | V4SH | O4SH | V6SH | O6SH | V8SH | O8SH | V2BLMT |
|---|---|---|---|---|---|---|---|---|---|
| V2SH | 1.00000 | 0.38280 | 0.67194 | 0.72200 | 0.64204 | 0.61679 | 0.73786 | 0.78320 | 0.88441 |
| V2SH | | 0.3092 | 0.0474 | 0.0281 | 0.0623 | 0.0769 | 0.0232 | 0.0125 | 0.0015 |
| O2SH | 0.38280 | 1.00000 | 0.02295 | 0.45034 | -0.19897 | -0.38525 | -0.14301 | -0.02118 | 0.01383 |
| O2SH | 0.3092 | | 0.9533 | 0.2238 | 0.6078 | 0.3059 | 0.7136 | 0.9569 | 0.9718 |
| V4SH | 0.67194 | 0.02295 | 1.00000 | 0.87049 | 0.74752 | 0.59619 | 0.92400 | 0.95629 | 0.84076 |
| V4SH | 0.0474 | 0.9533 | | 0.0023 | 0.0206 | 0.0902 | 0.0004 | <.0001 | 0.0045 |
| O4SH | 0.72200 | 0.45034 | 0.87049 | 1.00000 | 0.59920 | 0.33839 | 0.69832 | 0.80064 | 0.67449 |
| O4SH | 0.0281 | 0.2238 | 0.0023 | | 0.0882 | 0.3731 | 0.0364 | 0.0095 | 0.0463 |
| V6SH | 0.64204 | -0.19897 | 0.74752 | 0.59920 | 1.00000 | 0.92046 | 0.85839 | 0.85522 | 0.81205 |
| V6SH | 0.0623 | 0.6078 | 0.0206 | 0.0882 | | 0.0004 | 0.0031 | 0.0033 | 0.0078 |
| O6SH | 0.61679 | -0.38525 | 0.59619 | 0.33839 | 0.92046 | 1.00000 | 0.82889 | 0.77836 | 0.81852 |
| O6SH | 0.0769 | 0.3059 | 0.0902 | 0.3731 | 0.0004 | | 0.0057 | 0.0135 | 0.0070 |
| V8SH | 0.73786 | -0.14301 | 0.92400 | 0.69832 | 0.85839 | 0.82889 | 1.00000 | 0.98513 | 0.95187 |
| V8SH | 0.0232 | 0.7136 | 0.0004 | 0.0364 | 0.0031 | 0.0057 | | <.0001 | <.0001 |
| O8SH | 0.78320 | -0.02118 | 0.95629 | 0.80064 | 0.85522 | 0.77836 | 0.98513 | 1.00000 | 0.94336 |
| O8SH | 0.0125 | 0.9569 | <.0001 | 0.0095 | 0.0033 | 0.0135 | <.0001 | | 0.0001 |
| V2BLMT | 0.88441 | 0.01383 | 0.84076 | 0.67449 | 0.81205 | 0.81852 | 0.95187 | 0.94336 | 1.00000 |
| V2BLMT | 0.0015 | 0.9718 | 0.0045 | 0.0463 | 0.0078 | 0.0070 | <.0001 | 0.0001 | |
| O2BLMT | 0.58057 | -0.08882 | 0.82041 | 0.74014 | 0.83534 | 0.70679 | 0.83712 | 0.85281 | 0.72081 |
| O2BLMT | 0.1012 | 0.8202 | 0.0067 | 0.0226 | 0.0051 | 0.0333 | 0.0049 | 0.0035 | 0.0284 |
| V4BLMT | -0.18104 | -0.73595 | 0.42599 | 0.07837 | 0.28184 | 0.31801 | 0.42873 | 0.36061 | 0.18205 |
| V4BLMT | 0.6411 | 0.0238 | 0.2529 | 0.8412 | 0.4625 | 0.4043 | 0.2496 | 0.3464 | 0.6392 |
| O4BLMT | -0.32965 | -0.70896 | 0.25040 | -0.14735 | 0.01626 | 0.11762 | 0.27756 | 0.17024 | 0.06839 |
| O4BLMT | 0.3863 | 0.0325 | 0.5158 | 0.7052 | 0.9669 | 0.7631 | 0.4696 | 0.6615 | 0.8612 |
| V6BLMT | 0.70266 | -0.15348 | 0.95038 | 0.73130 | 0.86414 | 0.80524 | 0.99156 | 0.98756 | 0.92102 |
| V6BLMT | 0.0348 | 0.6934 | <.0001 | 0.0252 | 0.0027 | 0.0088 | <.0001 | <.0001 | 0.0004 |
| O6BLMT | 0.79227 | -0.04503 | 0.92122 | 0.73225 | 0.85896 | 0.81813 | 0.99053 | 0.98900 | 0.96845 |
| O6BLMT | 0.0109 | 0.9084 | 0.0004 | 0.0249 | 0.0030 | 0.0070 | <.0001 | <.0001 | <.0001 |
| V6ND | 0.68503 | -0.23268 | 0.92917 | 0.69061 | 0.88619 | 0.83576 | 0.98409 | 0.97065 | 0.91628 |
| V6ND | 0.0417 | 0.5469 | 0.0003 | 0.0394 | 0.0015 | 0.0050 | <.0001 | <.0001 | 0.0005 |

| Variable | O2BLMT | V4BLMT | O4BLMT | V6BLMT | O6BLMT | V6ND | O6ND | V8ND | O8ND |
|---|---|---|---|---|---|---|---|---|---|
| V2SH | 0.58057 | -0.18104 | -0.32965 | 0.70266 | 0.79227 | 0.68503 | 0.67952 | 0.94122 | 0.76479 |
| V2SH | 0.1012 | 0.6411 | 0.3863 | 0.0348 | 0.0109 | 0.0417 | 0.0441 | 0.0002 | 0.0164 |
| O2SH | -0.08882 | -0.73595 | -0.70896 | -0.15348 | -0.04503 | -0.23268 | -0.18894 | 0.45332 | 0.61972 |
| O2SH | 0.8202 | 0.0238 | 0.0325 | 0.6934 | 0.9084 | 0.5469 | 0.6264 | 0.2204 | 0.0751 |
| V4SH | 0.82041 | 0.42599 | 0.25040 | 0.95038 | 0.92122 | 0.92917 | 0.94225 | 0.53547 | 0.24279 |
| V4SH | 0.0067 | 0.2529 | 0.5158 | <.0001 | 0.0004 | 0.0003 | 0.0001 | 0.1373 | 0.5291 |
| O4SH | 0.74014 | 0.07837 | -0.14735 | 0.73130 | 0.73225 | 0.69061 | 0.73783 | 0.60156 | 0.40457 |
| O4SH | 0.0226 | 0.8412 | 0.7052 | 0.0252 | 0.0249 | 0.0394 | 0.0232 | 0.0866 | 0.2801 |
| V6SH | 0.83534 | 0.28184 | 0.01626 | 0.86414 | 0.85896 | 0.88619 | 0.89287 | 0.40912 | 0.07658 |
| V6SH | 0.0051 | 0.4625 | 0.9669 | 0.0027 | 0.0030 | 0.0015 | 0.0012 | 0.2742 | 0.8447 |
| O6SH | 0.70679 | 0.31801 | 0.11762 | 0.80524 | 0.81813 | 0.83576 | 0.79944 | 0.43947 | 0.13238 |
| O6SH | 0.0333 | 0.4043 | 0.7631 | 0.0088 | 0.0070 | 0.0050 | 0.0097 | 0.2366 | 0.7342 |
| V8SH | 0.83712 | 0.42873 | 0.27756 | 0.99156 | 0.99053 | 0.98409 | 0.96340 | 0.61069 | 0.30530 |
| V8SH | 0.0049 | 0.2496 | 0.4696 | <.0001 | <.0001 | <.0001 | <.0001 | 0.0807 | 0.4243 |
| O8SH | 0.85281 | 0.36061 | 0.17024 | 0.98756 | 0.98900 | 0.97065 | 0.96300 | 0.64522 | 0.34464 |
| O8SH | 0.0035 | 0.3404 | 0.6615 | <.0001 | <.0001 | <.0001 | <.0001 | 0.0606 | 0.3637 |
| V2BLMT | 0.72081 | 0.18205 | 0.06839 | 0.92102 | 0.96045 | 0.91628 | 0.89615 | 0.79361 | 0.52734 |
| V2BLMT | 0.0284 | 0.6392 | 0.8612 | 0.0004 | <.0001 | 0.0005 | 0.0011 | 0.0107 | 0.1446 |
| O2BLMT | 1.00000 | 0.51157 | 0.22636 | 0.84409 | 0.79946 | 0.84525 | 0.84375 | 0.42613 | 0.15445 |
| O2BLMT | | 0.1592 | 0.5581 | 0.0042 | 0.0097 | 0.0041 | 0.0042 | 0.2528 | 0.6915 |
| V4BLMT | 0.51157 | 1.00000 | 0.89451 | 0.45867 | 0.31300 | 0.49176 | 0.45649 | -0.23154 | -0.42642 |
| V4BLMT | 0.1592 | | 0.0011 | 0.2143 | 0.4122 | 0.1788 | 0.2168 | 0.5489 | 0.2524 |
| O4BLMT | 0.22636 | 0.89451 | 1.00000 | 0.27811 | 0.16277 | 0.30487 | 0.25833 | -0.27269 | -0.36494 |
| O4BLMT | 0.5581 | 0.0011 | | 0.4687 | 0.6756 | 0.4250 | 0.5021 | 0.4778 | 0.3342 |
| V6BLMT | 0.84409 | 0.45867 | 0.27811 | 1.00000 | 0.98322 | 0.99025 | 0.97667 | 0.54712 | 0.22324 |
| V6BLMT | 0.0042 | 0.2143 | 0.4687 | | <.0001 | <.0001 | <.0001 | 0.1274 | 0.5637 |
| O6BLMT | 0.79946 | 0.31300 | 0.16277 | 0.98322 | 1.00000 | 0.96610 | 0.94955 | 0.65847 | 0.36103 |
| O6BLMT | 0.0097 | 0.4122 | 0.6756 | <.0001 | | <.0001 | <.0001 | 0.0538 | 0.3398 |
| V6ND | 0.84525 | 0.49176 | 0.30487 | 0.99025 | 0.96610 | 1.00000 | 0.99133 | 0.52490 | 0.18661 |
| V6ND | 0.0041 | 0.1788 | 0.4250 | <.0001 | <.0001 | | <.0001 | 0.1468 | 0.6307 |

| Variable | V2SH | O2SH | V4SH | O4SH | V6SH | O6SH | V8SH | O8SH | V2BLMT |
|---|---|---|---|---|---|---|---|---|---|
| O6ND | 0.67952 | -0.18894 | 0.94225 | 0.73783 | 0.89287 | 0.79944 | 0.96340 | 0.96300 | 0.89615 |
| O6ND | 0.0441 | 0.6264 | 0.0001 | 0.0232 | 0.0012 | 0.0097 | <.0001 | <.0001 | 0.0011 |
| V8ND | 0.94122 | 0.45332 | 0.53547 | 0.60156 | 0.40912 | 0.43947 | 0.61069 | 0.64522 | 0.79361 |
| V8ND | 0.0002 | 0.2204 | 0.1373 | 0.0866 | 0.2742 | 0.2366 | 0.0807 | 0.0606 | 0.0107 |
| O8ND | 0.76479 | 0.61972 | 0.24279 | 0.40457 | 0.07658 | 0.13238 | 0.30530 | 0.34464 | 0.52734 |
| O8ND | 0.0164 | 0.0751 | 0.5291 | 0.2801 | 0.8447 | 0.7342 | 0.4243 | 0.3637 | 0.1446 |

CLUSTER = 7

| Variable | O2BLMT | V4BLMT | O4BLMT | V6BLMT | O6BLMT | V6ND | O6ND | V8ND | O8ND |
|---|---|---|---|---|---|---|---|---|---|
| O6ND | 0.84375 | 0.45649 | 0.25833 | 0.97667 | 0.94955 | 0.99133 | 1.00000 | 0.50812 | 0.16718 |
| O6ND | 0.0042 | 0.2168 | 0.5021 | <.0001 | <.0001 | <.0001 | | 0.1625 | 0.6673 |
| V8ND | 0.42613 | -0.23154 | -0.27269 | 0.54712 | 0.65847 | 0.52490 | 0.50812 | 1.00000 | 0.92556 |
| V8ND | 0.2528 | 0.5489 | 0.4778 | 0.1274 | 0.0538 | 0.1468 | 0.1625 | | 0.0003 |
| O8ND | 0.15445 | -0.42642 | -0.36494 | 0.22324 | 0.36103 | 0.18661 | 0.16718 | 0.92556 | 1.00000 |
| O8ND | 0.6915 | 0.2524 | 0.3342 | 0.5637 | 0.3398 | 0.6307 | 0.6673 | 0.0003 | |