FINAL REPORT

Leveraging Connected Vehicles to Enhance Traffic Responsive Traffic Signal Control

Date: May 2019

Prepared by:

Shrikant Fulari
Montasir Abbas

Virginia Polytechnic Institute
and State University
Blacksburg, VA 24061

Behrouz Salahshour, Mecit Cetin,

Transportation Research Institute
Old Dominion University
135 Kaufman Hall
Norfolk, VA, 23529

Wael Zatar, and Andrew P. Nichols
Marshall University
One John Marshall Drive
Huntington, WV 25537

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Leveraging Connected Vehicles to Enhance Traffic Responsive Traffic Signal Control		**5. Report Date** May 2019 **6. Performing Organization Code**
7. Author(s) Shrikant Fulari, Montasir Abbas, Behrouz Salahshour, Mecit Cetin, Wael Zatar, and Andrew P. Nichols		**8. Performing Organization Report No.**
9. Performing Organization Name and Address Virginia Polytechnic Institute and State University, Old Dominion University, and Marshall University		**10. Work Unit No. (TRAIS** **11. Contract or Grant No.**
12. Sponsoring Agency Name and Address US Department of Transportation Office of the Secretary-Research UTC Program, RDT-30 1200 New Jersey Ave., SE Washington, DC 20590		**13. Type of Report and Period Covered** Final **14. Sponsoring Agency Code**

15. Supplementary Notes

16. Abstract

For traffic signal control, Time of Day (TOD) mode of operations is widely deployed in practice for selecting a signal timing plan. However, TOD mode in not effective in adapting to variations in traffic conditions, such as special events and holidays, incidents, etc. Several research studies have reported the potential of Traffic Responsive Control operation or Traffic Responsive Plan Selection (TRPS) in reducing delays and the number of stops. For successful implementation of TRPS, accurate traffic state estimation is essential. The current study in this direction investigates a methodology for traffic state estimation for a corridor in Morgantown, WV, by using system detector data and connected vehicles (CV) data. Data from CVs form the basis to estimate queue lengths at signalized intersection approaches. While using data from multiple sources, a single measure in terms of three plan selection parameter was obtained, based on which discriminant functions were developed to classify the observations into states. Based on k-means clustering, similar traffic states were grouped together and a new set of states were suggested in place of the original states for which up to 93% classification accuracy was obtained. Overall, it was demonstrated that queue length data can be a valuable source of information for traffic state estimation that is needed for implementing the TRPS framework.

17. Key Words	18. Distribution Statement
Traffic signal control, traffic responsive plan selection, simulation, queue length estimation.	No restrictions. This document is available from the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	34	

Acknowledgements

The authors would like to thank Mid-Atlantic Transportation Sustainability Center – Region 3 University Transportation Center (MATS UTC) for funding this project. Dr. Andrew Nichols contributed to the development the research plan and assisted with the initial phase of this project before leaving Marshall University. The author and coauthors are thankful to Dr. Nichols for his contributions to the initial phase of this project.

Disclaimer

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

In traffic signal control, compared to the Time of Day (TOD) operation, the practice of using Traffic Responsive Plan Selection (TRPS) mode of operation is limited due to the simple and easy configuration of TOD mechanism. However, traffic patterns do change regularly on an hourly, daily or monthly basis, and TRPS mode of operation provides the flexibility of accommodating those traffic conditions by selecting the best suitable plan to minimize delay and stops and hence improve the overall system efficiency. TRPS mode also does not need timing plans to be updated as frequently as the TOD mode does, since the variations in traffic are incorporated into the mechanism while developing the plans. The TOD mode on the other hand does not offer this flexibility as the timing plans are pre-selected based on the time of the day. It is assumed that the traffic patterns are recurrent in time based on weekdays/weekends, etc. Hence variations in traffic conditions, such as special events/holidays, construction/work zone detours, random change in traffic patterns etc., do result in large delays and stops in the network under the TOD mode. The TOD timing plans have to be updated frequently based on the changes in traffic patterns making the procedure time consuming and labor intensive.

Past research have shown great potential and advantages of TRPS mode over TOD mode. Accurate sensing of traffic and performing accurate traffic state estimation is vital for the implementation of TRPS. The input data for state estimation is obtained mainly through the system detectors, while additional real-time traffic data can also be obtained through several other sources such as Bluetooth devices, connected/autonomous vehicles, etc. In the current study, an attempt is made to perform traffic state estimation while using system detector data and connected vehicles data (in the form of queue lengths obtained from simulation) from a given network.

Counts (volume) and occupancy (percentage of time the detector was occupied by vehicles) data mainly collected from system detectors in the network are currently widely used in practice to measure/analyze the traffic conditions. The emergence of Connected Vehicles (CV)/Autonomous Vehicles (AV) provides a new opportunity for obtaining real-time traffic data as these vehicles can transmit valuable information such as speed, position (based on which queue lengths can be obtained), etc., and these can further be used to estimate the current state of the system. While obtaining traffic data from multiple sources is now possible, there has to be a methodology of incorporating and processing such valuable data and deriving a single measure to identify different traffic states/conditions.

In the current study, a corridor from Morgantown, WV was selected for analysis. Traffic data from multiple sources were collected and analyzed to identify different traffic states. The simulations for the network were conducted using VISSIM 9, and the analysis was performed in MATLAB and Statistical Analysis Software (SAS). A discussion regarding past research in traffic state estimation and traffic signal control is provided in the next section.

BACKGROUND

Several data sources are currently available and in practice to obtain traffic data. This array includes field loop detectors, video cameras, infrared detectors, radar based detectors, Bluetooth sensors, probe vehicles equipped with Global Positioning System (GPS),

Connected/Autonomous vehicles, advanced communication systems such as Vehicle to Vehicle (V2V), Vehicle to infrastructure (V2I) etc. Hence, collection of traffic data through an automated system over long durations is now practically feasible. Such type of traffic data can be used for several traffic related applications such as traffic state estimation, traffic control and management/ traffic operations, Intelligent Transportation System (ITS) etc. Traffic control is one of the vital areas where traffic signals are widely used to regulate traffic. Two types of methodologies are widely considered in signal control, namely TOD mode and TRPS mode. Several studies have shown great potential of TRPS mode over TOD mode through research. TRPS mode heavily relies on accurate traffic sensing and the design of timing plans based on the thresholds developed/traffic states defined. Hence, accurate traffic state estimation can be considered as one of the primary stages of this methodology. Some studies that have reported different approaches in state estimation are discussed below.

Box et al. [1] discussed a methodology for instantaneous state estimation of an urban traffic network where data from multiple sensors, including wireless devices and inductive loops was used. The state was considered to be an estimate of the current distribution of vehicles in the network and their instantaneous speeds and this was obtained using an Extended Kalman Filter (EKF) approach. The method had better performance while estimating the number of vehicles however the performance was reduced while obtaining average speed of vehicles. Liu and Di [2] reported a study on traffic density estimation using fixed point data and GPS speed data on the signalized arterials. Srinivasan et al. [3] reported a study on the use of Neural Networks for real time traffic signal control. The study reported a multi agent system approach to develop distributed unsupervised traffic responsive signal control models by considering the local traffic signal controller for one intersection as an agent. The simulation study showed significant reduction in delays and mean stoppage time. Khan et al. [4] reported a study for real time traffic state estimation with the help of connected vehicles. The study focused on increasing the real-time roadway traffic condition assessment accuracy by using connected vehicle technology and artificial intelligence. On comparison of the Level of Service estimation with the Caltrans Performance Measurement System (PeMS), the performance of this study was observed to be better. Mannini et al. [5] reported a study on a methodology to estimate route travel time based on historical and real-time data obtained from multiple sources that were obtained through advanced monitoring systems. The study used a second order macroscopic traffic flow model with an Extended Kalman Filter (EKF) approach with a data fusion technique while using simulated data.

While the above studies have focused on the direct application of state estimates such as density, travel time etc., these estimates cannot be directly incorporated for TRPS operation. For TRPS operation, the real time data from field system detectors such as counts, occupancy, queue lengths, average speed etc., are vital. Use of on field system detectors is one of the widely practiced methods of obtaining data for traffic signal control.

Abbas et al. [6] reported a study on the methodology for determining optimal traffic responsive plan selection control parameters. The study focused on developing optimal timing plans that are suitable for a wide range of traffic conditions and mapping the different traffic conditions to one of the available timing plans which are stored in traffic controllers. The study mainly used genetic algorithms and discriminant analysis in the framework and the data used was from

system detectors placed in the field. Abbas and Abdelaziz [7] reported a study on evaluation of traffic responsive control for an arterial network while considering the issues of unequal traffic distribution and large combination of traffic movements from multiple intersections. The study implemented a multi-objective optimization method to generate final timing plans and TRPS pattern matching parameters. Count and occupancy data obtained from system detectors in the network were used in the framework. Abbas and Sharma [8] reported a study where they proposed use of a multi-objective evolutionary algorithm for optimizing the TOD plan scheduling, and proposed a new measure of performance namely Degree of Detachment (DOD) for providing a clustering mechanism of traffic patterns. The study made use of traffic data collected from system detectors. In another study, Abbas and Sharma [9] reported a new robust methodology for the selection of TRPS optimal parameters and thresholds by using Bayesian-based discriminant analysis. While using field data collected through system detectors, this study reported a 100% classification accuracy using this approach. Abbas et al. [10] reported a methodology for TRPS operation using multi-objective evolutionary algorithm and supervised discriminant analysis. The study developed nine timing plans to be used with the TRPS mode and performed the tests with simulated data, and reported a possibility of 53% savings in delay and 16% savings in stops in comparison to TOD mode of operation. A study by Sharma [11] discussed methodology for determination of traffic responsive plan selection factors and thresholds using Artificial Neural Networks (ANN). The study used k-means clustering for identifying demand states and further determination of TRPS weights and thresholds was performed using ANN. The study used data from system detectors in the analysis.

From the above studies it can be observed that system detectors placed in the field are widely used for sensing the traffic, developing timing plans based on different traffic states and then for real time implementation of TRPS operation. As discussed earlier, with the advent of data collection technologies, real time traffic data can now be collected from several sources. The challenge remains to utilize this data into the TRPS development and operational framework. The current study in this direction focused on using real time data from mobile vehicles (as CV/AV) that would soon occupy the traffic stream. Data from these vehicles can be vital in providing real time estimate of the traffic quality and several other measures. Queue length is one such variable that can provide vital information regarding the quality of traffic at the location where they are obtained. Several studies have reported estimation of queue lengths using the data obtained from probe or connected vehicles (Li et al. [12], Badillo et al. [13]), and reported their applications in adaptive signal control (Tiaprasert [14]), developing measures of effectiveness for determining traffic conditions on urban signalized arterials for real time applications (Argote [15]) etc. Use of queue lengths hence can be very viral in numerous traffic applications.

However, not many studies have reported the fusion of such system detector/stationary sensor data and mobile vehicle data into the development and implementation of TRPS framework. In this study, counts and occupancy data from system detectors in combination with queue lengths obtained through VISSIM simulation is used to provide an estimate of traffic state. The queue length data can be assumed to be coming from connected vehicles in the future. Inclusion of such data from multiple sources can provide us with enhanced reliability regarding the traffic state and can be valuable for TRPS development and operation.

A discussion regarding the study site and methodology is provided in subsequent sections.

STUDY SITE AND DATA COLLECTION

Corridor description

A corridor in Morgantown, WV was selected for the analysis in this study. The selected arterial consisted of five signalized intersections. A google maps image of the corridor and the intersections (circled and numbered for reference) are show in Figure *1*, and the details are provided in Table *1*. A VISSIM network of the selected corridor was provided for the analysis and the reference number of the signal controllers associated with each of the five intersections are also provided in Table *1*.

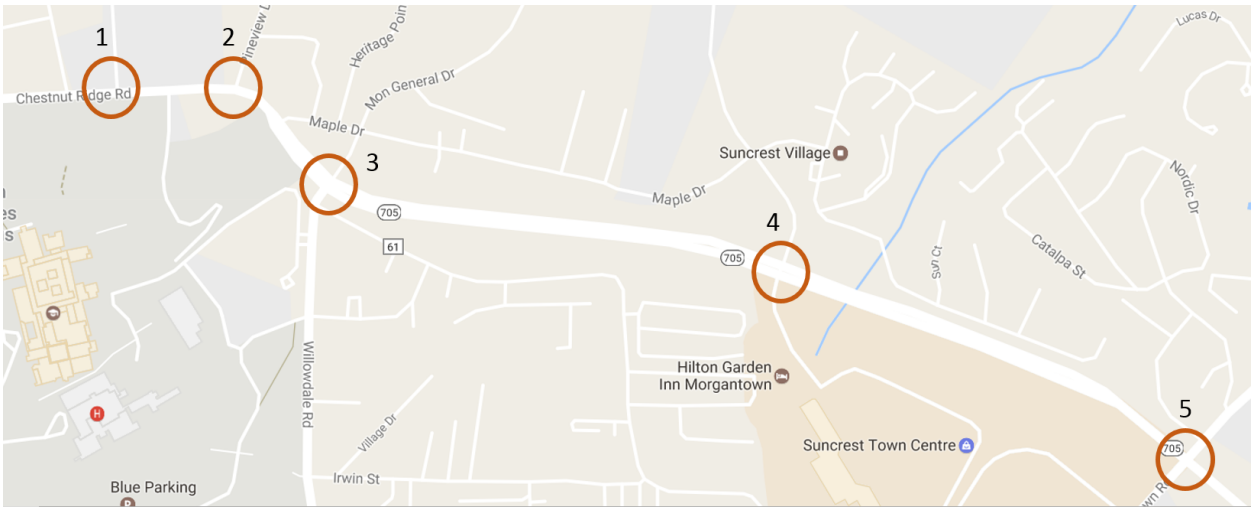

Figure 1 Study corridor and intersections (Source: Google maps)

Table 1 Details about the intersections

Intersection reference number	Name	Coordinates	VISSIM Signal controller number
1	Chestnut Ridge road and North elementary school road	39.658011, -79.956882	1010
2	Chestnut Ridge road and Pineview drive	39.658061, -79.954670	1011
3	Chestnut Ridge road and Willowdale road	39.656684, -79.952922	1012
4	WV 705 and WVU Research Park	39.655434, -79.944440	1013
5	WV 705 and Stewartstown road	39.652665, -79.936807	1014

Figure *2* and Figure *3* provide details regarding the placement and identification of system detectors in the network covering all the intersections, and the queue counters placed close to the intersections.

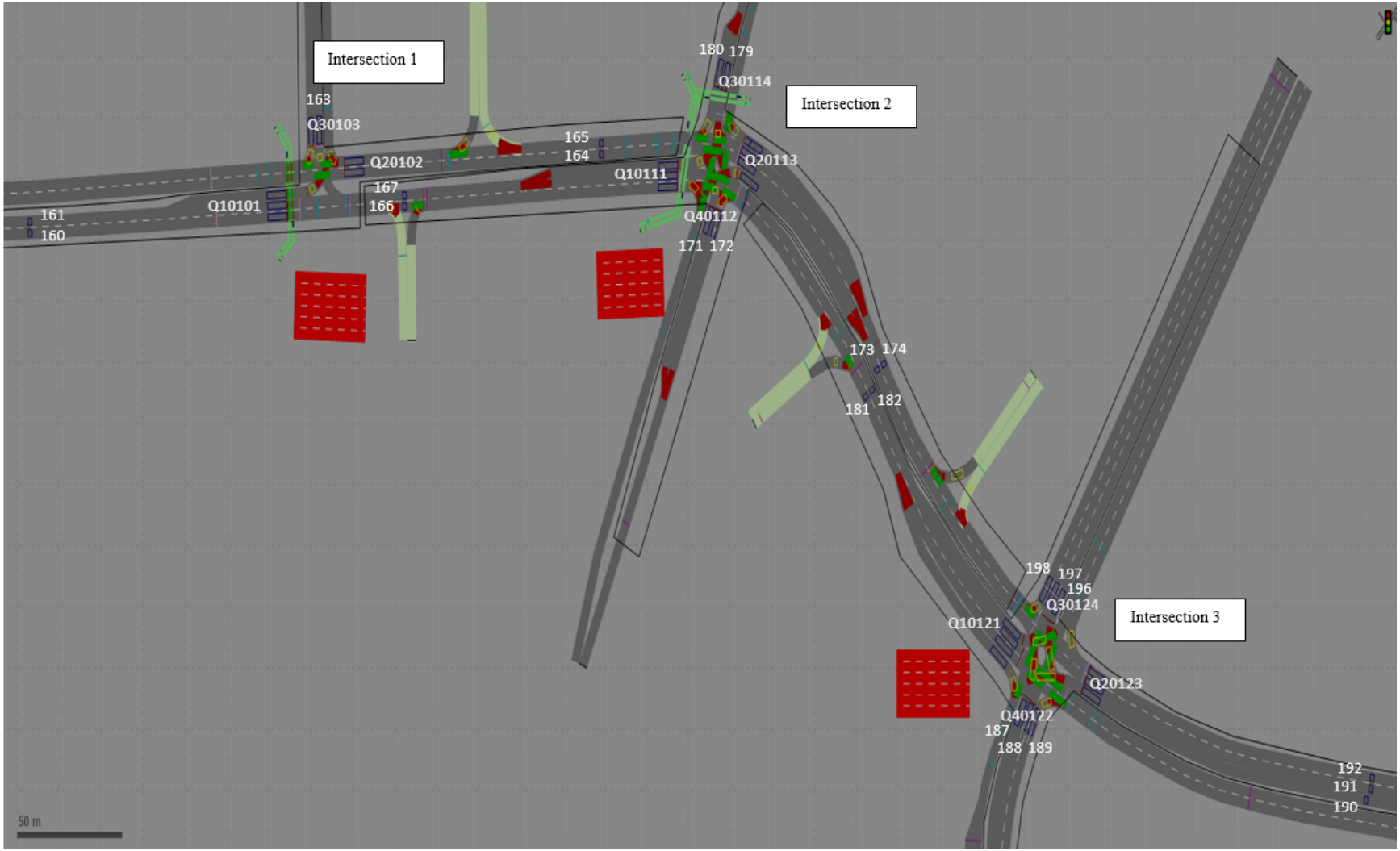

Figure 2 VISSIM network consisting of first three intersections, along with the location and identification of system detectors and queue counters

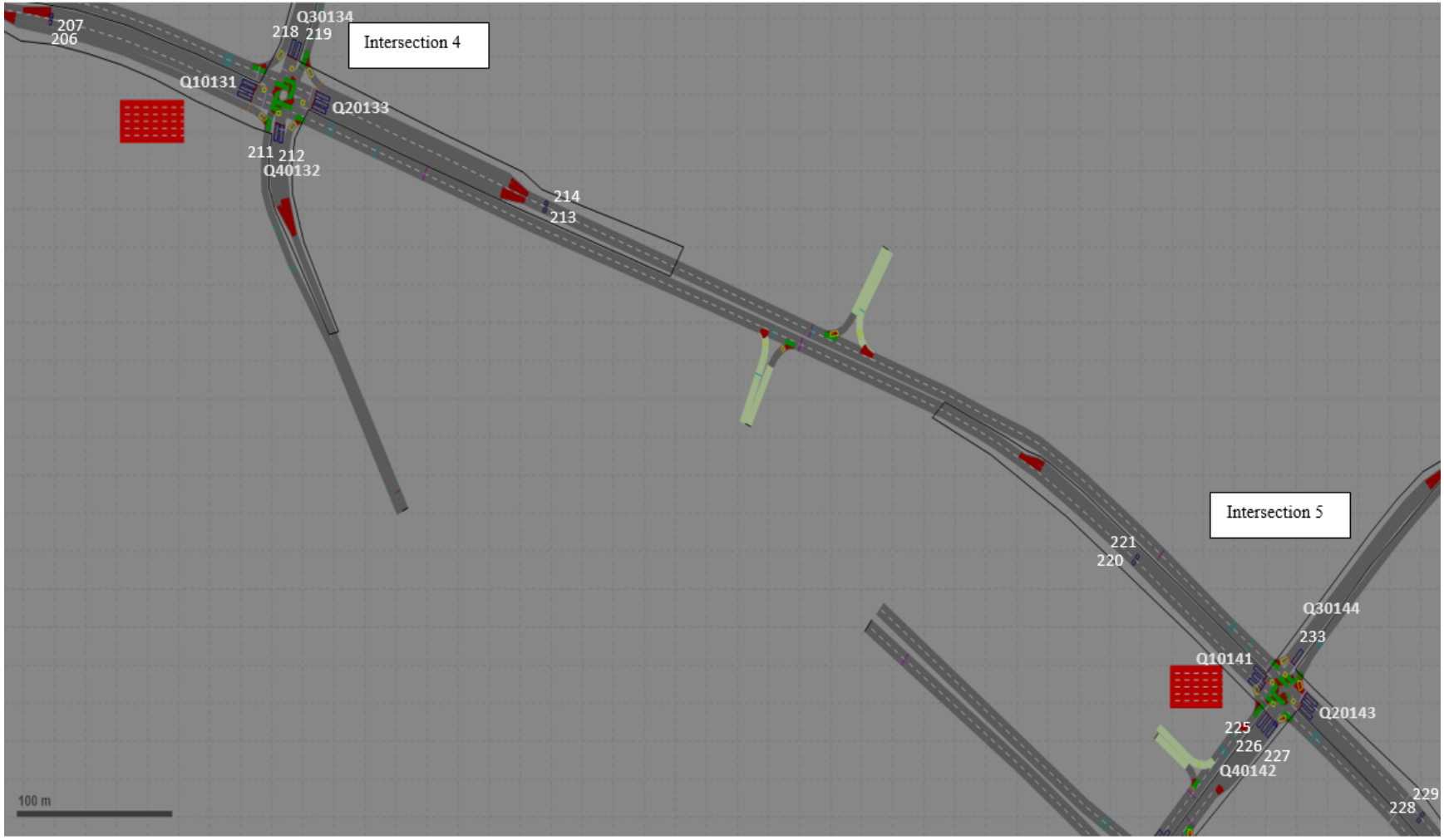

Figure 3 VISSIM network consisting of remaining two intersections, along with the location and identification of system detectors and queue counters

Signal controller information
The five Ring Barrier Control (RBC) signal controllers associated with five intersections in the corridor were set with the provided 10 different timing plans. The side street phase movements were set on detector actuation mode, meaning the side street would be served green only when a detector call was placed, else the main street movement would have the green.

Data collection from VISSIM
The total duration of this analysis was 15 hours (54000 seconds). Counts, occupancy, and queue lengths were the main variables used in this study for state estimation. Count and occupancy were obtained from 40 system detectors placed all over the network covering side streets and main arterial links as shown in Figure *2* and Figure *3* (identified with numbers: 160, 161 etc.). Queue lengths were obtained from the queue counters placed in the network as shown in Figure *2* and Figure *3* (identified as: Q10101, Q20102 etc.). Counts, occupancy and average queue lengths were collected at every 10-minute interval. The simulations were run for all the 10 different timing plans. These timing plans for each intersection are provided in the tables in Appendix I.

OBJECTIVES AND SCOPE

The key objectives identified for the study are:

1. Develop a framework in MATLAB for data collection from system detectors and queue counters in VISSIM network, and obtain count, occupancy and queue length data.
2. Perform a canonical discriminant analysis to identify weights associated with each of the variables used for obtaining three different plan selection (PS) parameters (namely cycle, offset, and split).
3. Use the weights and input data, compute all three PS parameters associated with each observation, and further obtain discriminant functions for each state
4. Determine the thresholds of PS parameter to switch from one state to another based on each PS parameter.
5. Perform k-means clustering to identify similar traffic states that can be grouped together
6. Perform a comparison analysis of the classification of states for data obtained at different penetration rates.

The data collection, estimation and analysis in this study is performed for the corridor consisting of Chestnut Ridge road and WV 705 only, which include five signalized intersections. The current study would focus only on the traffic state estimation part.

METHODOLOGY AND RESULTS

The adopted methodology could be briefly described as: Obtaining count, occupancy and queue lengths from the network, obtaining weights for the selected variables, computing the plan selection parameter, obtaining discriminant functions for each state, and then identifying thresholds for switching between states.

Three PS parameters were estimated in this study at the cycle level, offset level, and split level. To calculate the cycle level PS parameter, the detectors located at critical locations were used. To calculate the offset level PS parameter, the detectors placed on arterials in the inbound and outbound directions were used. To calculate the split level PS parameter, the detectors placed on the non-arterials/side streets were used. Similar selection was applied to the queue counters in the network. The selected detectors at each level and the associated weights are represented in Table *5*, Table *6*, and Table *7*.

An important consideration in the study was to analyze the estimation accuracy at different penetration rates of the connected vehicles in the network. As queue lengths collected in this study through simulation represent the queue lengths obtained from connected vehicles in the real network, to obtain data at different penetration rates, results from Li et al., [12] for estimation of queue lengths at different penetration rates were used. These were used to introduce errors/generate perturbations in the ground truth queue lengths in the network. The queue lengths obtained through queue counters in VISSIM were considered as ground truth queue lengths as these are obtained from all the vehicles in the network, meaning data obtained at 100% penetration rate. Table *2* represents the error rates (Mean absolute percentage error: MAPE) used from Li et al., [12] in the current study to be introduced in the ground truth queue length data.

Table 2 Error rates for queue lengths at different penetration rates

Penetration rate (%)	Error rate (MAPE %)
90	4.29
80	6.35
70	11.35
60	14.26
50	17.27
40	24.95
30	29.80
20	42.15
10	60.82

Another important consideration was the selection of timing plans from the provided 10 different timing plans (Appendix I) for the simulations. In order to select data from the most suitable timing plans, initially the simulations were run for all the 10 timing plans for 15 hours (covering all states). From the simulations, the queuing delays were obtained from all the queue counters at five intersections, and sum of delays from all the queue counters was obtained for each hour (state). From these results, the optimal timing plans (based on the least delay) for each state as well as overall delays for all states are determined. Table *3* presents the computed delays.

Table 3 Summary of sum of queuing delays (seconds) for each state and timing plan combination

Timing plan ▶	1	2	3	4	5	6	7	8	9	10

State										
1	3072.2	3280.9	**3045.8**	3107.9	3061.9	3156.3	3358.5	3498.8	4132.3	4342.7
2	**5334.1**	5379.9	5691.1	5807.3	5770.2	5733.9	5920.2	6211.1	7023.6	6749.0
3	9599.9	**8463.5**	10278.3	10318.6	9254.5	10819.3	10647.3	9789.1	12539.4	11654.2
4	16296.3	14549.1	15930.1	15265.2	**13962.4**	16050.8	15645.3	14817.5	19799.7	17189.1
5	14840.8	13886.5	14485.8	13978.7	**12904.7**	15332.3	14767.5	13456.1	17997.6	15624.5
6	11430.7	**11157.0**	12019.8	11951.0	12068.1	12785.9	12135.6	11938.3	14501.1	13590.3
7	11579.9	**11092.0**	11594.3	12000.5	11775.9	12049.7	12160.1	12140.0	14241.3	13881.6
8	**11173.4**	11988.9	11800.6	11649.6	11782.8	12663.8	12379.9	12633.2	15146.4	14897.1
9	**11691.6**	11989.8	12174.0	11978.6	11989.8	12448.3	11752.2	11813.6	15509.8	15030.9
10	12086.7	12002.9	12863.3	12126.8	**11967.1**	13214.3	12892.5	13081.9	15390.3	15140.7
11	**11546.1**	12041.4	12245.0	12033.8	11882.9	13200.1	12890.7	12595.4	16171.6	15509.6
12	15862.4	17054.0	15901.4	16866.3	17727.3	**15191.4**	19692.6	16944.0	20589.3	26024.5
13	**12189.0**	14974.1	16399.0	14370.4	14363.2	15115.6	16400.9	15268.8	21611.2	23544.5
14	**14147.5**	14339.3	16644.9	15175.9	15380.7	16236.8	16887.6	15987.1	23351.2	19519.6
15	10432.8	**9527.2**	9866.6	11070.5	10200.1	9793.1	10741.3	11333.4	13264.9	12292.4
Total	**171283.3**	**171726.4**	**180940.1**	**177701.0**	**174091.7**	183791.8	188272.2	181508.3	231269.5	224990.7

From Table *3*, it can be observed that timing plans 1, 2, 5, 4 and 3 were the five timing plans that had low total delays. Fourteen out of fifteen states had one of these five timing plans as their optimal timing plan (respective optimal timing plan is highlighted in the table). State 12 had timing plan 6 as the optimal, but since timing plan 6 had a high overall total delay, it was not included in the analysis. Hence the final analysis included timing plans 1, 2, 5, 4 and 3.

Weights associated with the selected variables

As mentioned earlier, data were obtained from a total of 40 system detectors and 19 queue counters. Each detector provides count and occupancy value, hence resulting in 80 different variables. Adding the queue variable from 19 queue counters results in a set of 99 different explanatory variables for identifying a state. This study focused on estimating three PS parameters, namely at the cycle level, offset level and split level. Each of these parameters is calculated by using a set of/combination of system detectors as mentioned earlier. In order to have a strong discriminatory power to discriminate different states, each of the PS parameter can have different set of weights for the detectors used for computing them.

A canonical discriminant analysis was performed on these variables to identify the weights associated with each variable for each PS parameter. Canonical discriminant analysis is a dimensionality reduction technique that provides a best linear combination of the selected variables by associating them with canonical coefficients. This linear combination is aimed at providing best discrimination among different classes considered. This procedure was performed in Statistical Analysis Software (SAS) package. Based on the data set and the number of explanatory variables, SAS provides a set of canonical variables along with their canonical relation value and the F-statistic value to identify the best canonical variable and its coefficients as weights. Table *4* provides statistics for these variables.

Table 4 Statistics of canonical discriminant analysis

	Cycle level		Offset level		Split level	
Canonical variable	**Canonical correlation**	**F value**	**Canonical correlation**	**F value**	**Canonical correlation**	**F value**
1	**0.988148**	**13.35**	**0.969998**	**11.18**	**0.977803**	**13.09**
2	0.974247	10.86	0.938643	8.73	0.950694	9.88
3	0.970869	9.19	0.892373	7.08	0.889162	7.73
4	0.920239	7.56	0.831391	5.95	0.837774	6.51
5	0.903882	6.84	0.801373	5.22	0.801313	5.62
6	0.892697	6.19	0.757573	4.54	0.757672	4.85
7	0.870369	5.53	0.725813	3.97	0.742099	4.18
8	0.853809	4.95	0.690599	3.4	0.646616	3.41
9	0.825234	4.35	0.630088	2.84	0.609806	2.95
10	0.793293	3.79	0.566473	2.36	0.519499	2.46
11	0.775881	3.27	0.522335	1.96	0.470498	2.19
12	0.720728	2.59	0.450459	1.52	0.427138	1.97
13	0.589356	1.9	0.333566	1.11	0.389303	1.77
14	0.543114	1.7	0.301104	1.01	0.33984	1.54

From Table *4*, it can be observed that the 1st canonical variable shows the highest correlation and significance as compared to other variables, hence the coefficients associated with the first canonical variable were used as the weights in further analysis. The associated raw weights for each variable are represented in Table *5* at the cycle level, in Table *6* at the offset level and in Table *7* at the split level. The prefix 'C' indicates count and 'O' indicates Occupancy, and it associated number indicates the system detector number from the VISSIM network. Similarly, the prefix 'Q' to the number indicates the queue counter.

Table 5 Weights associated with variables for Cycle level PS parameter

Variable	Weight	Variable	Weight	Variable	Weight
C160	0.012823	**C221**	-0.01307	**O213**	0.081337
C161	0.022493	**C228**	-0.01493	**O214**	-0.30913
C164	-0.00819	**C229**	-0.0211	**O211**	-0.0174
C165	-0.02775	**C225**	0.006733	**O212**	-0.00266
C163	-0.00447	**C226**	0.137672	**O218**	-0.01652

C166	0.02277	**C227**	0.045314	**O219**	0.010017
C167	-0.0275	**C233**	-0.03215	**O220**	-0.04896
C173	-0.01021	**O160**	0.062666	**O221**	-0.00485
C174	0.014545	**O161**	-0.00601	**O228**	0.063536
C171	0.022436	**O164**	-0.04067	**O229**	0.001499
C172	-0.08179	**O165**	0.021055	**O225**	-0.00251
C179	-0.00646	**O163**	-0.03439	**O226**	-0.0412
C180	0.029782	**O166**	0.031304	**O227**	0.010071
C181	-0.01045	**O167**	-0.03863	**O233**	-0.01187
C182	0.017889	**O173**	0.056426	**Q10101**	-0.00547
C190	-0.02498	**O174**	-0.01084	**Q30103**	0.02355
C191	-0.0474	**O171**	0.036699	**Q20102**	0.003626
C192	-0.00442	**O172**	-0.02151	**Q10111**	0.031809
C187	0.194224	**O179**	-0.01496	**Q30114**	0.005809
C188	0.010445	**O180**	-0.0369	**Q20113**	0.001673
C189	0.043319	**O181**	-0.05087	**Q40112**	-0.00159
C196	-0.1112	**O182**	-0.01081	**Q10121**	0.017052
C197	-0.11233	**O190**	0.186809	**Q30124**	0.086474
C198	0.156814	**O191**	0.723447	**Q20123**	0.001809
C206	-0.03067	**O192**	-0.30666	**Q40122**	0.005281
C207	-0.00901	**O187**	-0.01022	**Q10131**	0.055175
C213	0.003324	**O188**	-0.01967	**Q30134**	0.112813
C214	0.000993	**O189**	0.057904	**Q20133**	0.041674
C211	0.063142	**O196**	-0.00138	**Q40132**	0.139214
C212	0.032425	**O197**	-0.00454	**Q10141**	0.001864
C218	-0.20455	**O198**	0.046534	**Q30144**	0.005244
C219	0.072282	**O206**	-0.07412	**Q20143**	-0.02887
C220	0.010385	**O207**	0.037479	**Q40142**	0.006214

Table 6 Weights associated with variables for Offset level PS parameter

Variable	Weight	Variable	Weight	Variable	Weight
C160	0.036604	**C221**	-0.02002	**O213**	0.399292
C161	0.010756	**C228**	0.019361	**O214**	-0.60059
C164	-0.0028	**C229**	-0.02235	**O220**	0.019982
C165	0.012174	**O160**	0.048599	**O221**	0.017423
C166	0.000969	**O161**	-0.02948	**O228**	-0.0173

C167	-0.02628	**O164**	-0.00019	**O229**	-0.04209
C173	0.010304	**O165**	-0.04741	**Q10101**	-0.00386
C174	0.022022	**O166**	-0.01942	**Q20102**	0.020244
C181	-0.00739	**O167**	0.022414	**Q10111**	0.009861
C182	0.013031	**O173**	0.022183	**Q20113**	-0.01313
C190	-0.04155	**O174**	-0.00284	**Q10121**	0.021723
C191	0.009709	**O181**	-0.05392	**Q20123**	-0.00254
C192	-0.01416	**O182**	0.013201	**Q10131**	0.05147
C206	0.02506	**O190**	0.26844	**Q20133**	-0.02834
C207	0.039839	**O191**	0.022152	**Q10141**	0.000892
C213	-0.02166	**O192**	-0.03672	**Q20143**	0.009029
C214	0.00286	**O206**	-0.54569		
C220	0.025369	**O207**	0.050929		

Table 7 Weights associated with variables for split level PS parameter

Variable	Weight	Variable	Weight	Variable	Weight
C163	-0.00524	**O187**	-0.00794	**O218**	0.050003
C171	-0.04895	**O188**	-0.01262	**O219**	-0.00399
C172	-0.07971	**O189**	0.104056	**O225**	0.003227
C179	-0.01977	**O196**	0.006131	**O226**	-0.03833
C180	-0.02567	**O197**	-0.00871	**O227**	0.069977
C187	0.137047	**O198**	0.030515	**O233**	-0.01876
C188	-0.01541	**O211**	-0.02237	**Q30103**	0.023213
C189	0.01793	**C226**	0.14158	**Q30114**	0.003808
C196	-0.10905	**C227**	0.011476	**Q40112**	0.02722
C197	-0.09553	**C233**	-0.00275	**Q30124**	0.24804
C198	0.140934	**O163**	-0.0086	**Q40122**	-0.013
C211	0.025725	**O171**	0.035341	**Q30134**	-0.29938
C212	-0.01125	**O172**	-0.01907	**Q40132**	0.308923
C218	-0.1858	**O179**	-0.03011	**Q30144**	0.009652
C219	0.051431	**O180**	-0.01295	**Q40142**	0.008849
C225	-0.02359	**O212**	0.012112		

Discriminant functions for each state

Discriminant functions developed for individual classes based on a parameter are mainly used to classify a future observation into one of the classes based on the same parameter. For a given observation, the class function that produces highest value is assigned as the class label for that observation. These functions can be tested on the known observation data and the accuracy of classification can be analyzed by comparing it with ground truth class labels.
In the current study, the discriminant functions were developed for each of the 15 states for each of the three PS parameters. These discriminant functions were developed by using the data at

100% penetration rate. This was done based on the PS parameter value and the state label for each observation. PS parameter was computed as the sum of product of each variable with its assigned final weight. The discriminant functions at cycle, offset and split level are tabulated in Table *8*.

Table 8 Discriminant functions for each state at each PS parameter level

	Cycle		Offset		Split	
State	**Constant**	**Coefficient of Cycle PS parameter**	**Constant**	**Coefficient of Offset PS parameter**	**Constant**	**Coefficient of Split PS parameter**
1	-1.45884	-1.70812	-0.57031	-1.068	-0.13413	-0.51794
2	-16.2572	-5.70213	-5.52892	-3.32533	-2.18527	-2.09059
3	-38.8449	-8.81418	-5.18146	-3.21915	-6.00138	-3.4645
4	-27.1253	-7.3655	-0.53715	-1.03648	-8.62279	-4.15278
5	-31.3072	-7.91292	-0.8176	-1.27875	-10.8204	-4.65196
6	-26.4875	-7.2784	-1.33077	1.63143	-12.5867	-5.01731
7	-10.1476	-4.50502	-5.87918	3.42905	-6.93857	-3.7252
8	-3.84802	-2.77417	-11.0033	4.69112	-2.86341	-2.39308
9	-1.69249	-1.83983	-10.6835	4.62243	-0.68006	-1.16624
10	-2.09822	-2.04852	-12.3665	4.97323	-0.05266	-0.32452
11	-10.5773	4.59941	-20.2313	6.36102	-6.44496	3.59025
12	-79.054	12.5741	-42.7522	9.24686	-49.1953	9.9192
13	-53.245	10.3194	-31.8044	7.97551	-43.6546	9.34394
14	-7.99422	3.99855	-21.8464	6.61006	-7.2785	3.81536
15	-0.48068	0.98049	-1.01149	1.42231	-0.43813	0.93609

Figure *3* shows a plot of discriminant functions at the cycle level. It can be observed from Figure *4* that that the discriminant functions intersect each other at a certain point which can be identified as a threshold for making transition from one to another state.

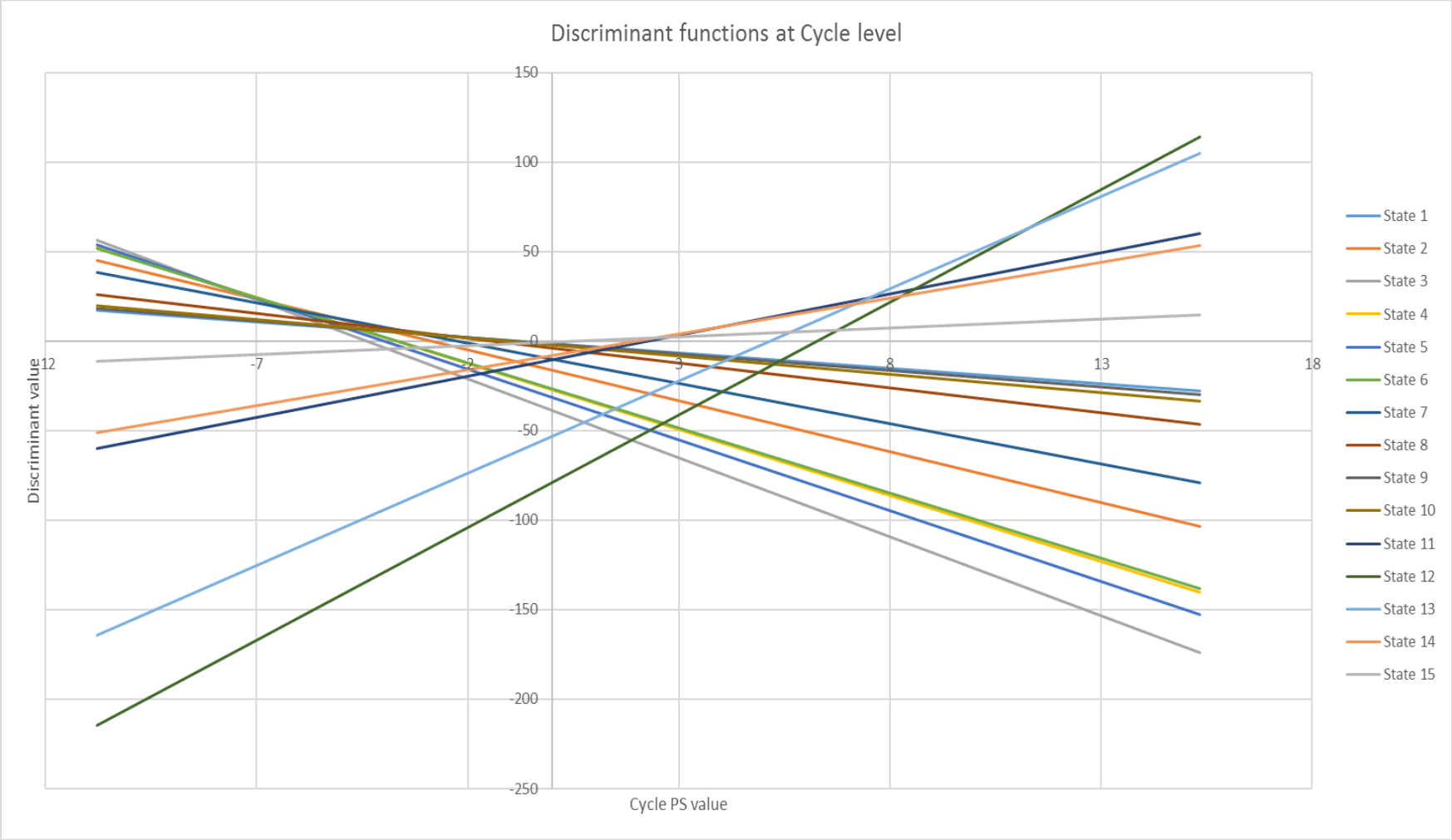

Figure 4 Plot of discriminant functions for Cycle PS parameter

Thresholds were hence identified based on the discriminant functions at each of the PS parameter level. The thresholds based on three variables for a single observation can help in accurately classifying the observation into one of the states. The threshold matrix was hence developed at each PS parameter level. These thresholds are tabulated in Table *9*, Table *10* and Table *11*for Cycle, Offset and Split level respectively.

Table 9 Threshold matrix for Cycle PS parameter

State	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		-3.71	-5.26	-4.54	-4.81	-4.49	-3.11	-2.24	-1.77	-1.88	1.45	5.43	4.31	1.15	-0.36
2	-3.71		-7.26	-6.53	-6.81	-6.49	-5.10	-4.24	-3.77	-3.88	-0.55	3.44	2.31	-0.85	-2.36
3	-5.26	-7.26		-8.09	-8.36	-8.05	-6.66	-5.79	-5.33	-5.43	-2.11	1.88	0.75	-2.41	-3.92
4	-4.54	-6.53	-8.09		-7.64	-7.32	-5.94	-5.07	-4.60	-4.71	-1.38	2.60	1.48	-1.68	-3.19
5	-4.81	-6.81	-8.36	-7.64		-7.60	-6.21	-5.34	-4.88	-4.98	-1.66	2.33	1.20	-1.96	-3.47
6	-4.49	-6.49	-8.05	-7.32	-7.60		-5.89	-5.03	-4.56	-4.66	-1.34	2.65	1.52	-1.64	-3.15
7	-3.11	-5.10	-6.66	-5.94	-6.21	-5.89		-3.64	-3.17	-3.28	0.05	4.03	2.91	-0.25	-1.76
8	-2.24	-4.24	-5.79	-5.07	-5.34	-5.03	-3.64		-2.31	-2.41	0.91	4.90	3.77	0.61	-0.90
9	-1.77	-3.77	-5.33	-4.60	-4.88	-4.56	-3.17	-2.31		-1.94	1.38	5.37	4.24	1.08	-0.43
10	-1.88	-3.88	-5.43	-4.71	-4.98	-4.66	-3.28	-2.41	-1.94		1.28	5.26	4.14	0.98	-0.53
11	1.45	-0.55	-2.11	-1.38	-1.66	-1.34	0.05	0.91	1.38	1.28		8.59	7.46	4.30	2.79
12	5.43	3.44	1.88	2.60	2.33	2.65	4.03	4.90	5.37	5.26	8.59		11.45	8.29	6.78
13	4.31	2.31	0.75	1.48	1.20	1.52	2.91	3.77	4.24	4.14	7.46	11.45		7.16	5.65
14	1.15	-0.85	-2.41	-1.68	-1.96	-1.64	-0.25	0.61	1.08	0.98	4.30	8.29	7.16		2.49
15	-0.36	-2.36	-3.92	-3.19	-3.47	-3.15	-1.76	-0.90	-0.43	-0.53	2.79	6.78	5.65	2.49	

Table 10 Threshold matrix for Offset PS parameter

State	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		-2.20	-2.14	-1.05	-1.17	0.28	1.18	1.81	1.78	1.95	2.65	4.09	3.45	2.77	0.18
2	-2.20		-3.27	-2.18	-2.30	-0.85	0.05	0.68	0.65	0.82	1.52	2.96	2.33	1.64	-0.95
3	-2.14	-3.27		-2.13	-2.25	-0.79	0.10	0.74	0.70	0.88	1.57	3.01	2.38	1.70	-0.90
4	-1.05	-2.18	-2.13		-1.16	0.30	1.20	1.83	1.79	1.97	2.66	4.11	3.47	2.79	0.19
5	-1.17	-2.30	-2.25	-1.16		0.18	1.08	1.71	1.67	1.85	2.54	3.98	3.35	2.67	0.07
6	0.28	-0.85	-0.79	0.30	0.18		2.53	3.16	3.13	3.30	4.00	5.44	4.80	4.12	1.53
7	1.18	0.05	0.10	1.20	1.08	2.53		4.06	4.03	4.20	4.90	6.34	5.70	5.02	2.43
8	1.81	0.68	0.74	1.83	1.71	3.16	4.06		4.66	4.83	5.53	6.97	6.33	5.65	3.06
9	1.78	0.65	0.70	1.79	1.67	3.13	4.03	4.66		4.80	5.49	6.93	6.30	5.62	3.02
10	1.95	0.82	0.88	1.97	1.85	3.30	4.20	4.83	4.80		5.67	7.11	6.47	5.79	3.20
11	2.65	1.52	1.57	2.66	2.54	4.00	4.90	5.53	5.49	5.67		7.80	7.17	6.49	3.89
12	4.09	2.96	3.01	4.11	3.98	5.44	6.34	6.97	6.93	7.11	7.80		8.61	7.93	5.33
13	3.45	2.33	2.38	3.47	3.35	4.80	5.70	6.33	6.30	6.47	7.17	8.61		7.29	4.70
14	2.77	1.64	1.70	2.79	2.67	4.12	5.02	5.65	5.62	5.79	6.49	7.93	7.29		4.02
15	0.18	-0.95	-0.90	0.19	0.07	1.53	2.43	3.06	3.02	3.20	3.89	5.33	4.70	4.02	

Table 11 Threshold matrix for Split PS parameter

State	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		-1.30	-1.99	-2.34	-2.58	-2.77	-2.12	-1.46	-0.84	-0.42	1.54	4.70	4.41	1.65	0.21
2	-1.30		-2.78	-3.12	-3.37	-3.55	-2.91	-2.24	-1.63	-1.21	0.75	3.91	3.63	0.86	-0.58
3	-1.99	-2.78		-3.81	-4.06	-4.24	-3.59	-2.93	-2.32	-1.89	0.06	3.23	2.94	0.18	-1.26
4	-2.34	-3.12	-3.81		-4.40	-4.59	-3.94	-3.27	-2.66	-2.24	-0.28	2.88	2.60	-0.17	-1.61
5	-2.58	-3.37	-4.06	-4.40		-4.83	-4.19	-3.52	-2.91	-2.49	-0.53	2.63	2.35	-0.42	-1.86
6	-2.77	-3.55	-4.24	-4.59	-4.83		-4.37	-3.71	-3.09	-2.67	-0.71	2.45	2.16	-0.60	-2.04
7	-2.12	-2.91	-3.59	-3.94	-4.19	-4.37		-3.06	-2.45	-2.02	-0.07	3.10	2.81	0.05	-1.39
8	-1.46	-2.24	-2.93	-3.27	-3.52	-3.71	-3.06		-1.78	-1.36	0.60	3.76	3.48	0.71	-0.73
9	-0.84	-1.63	-2.32	-2.66	-2.91	-3.09	-2.45	-1.78		-0.75	1.21	4.38	4.09	1.32	-0.12
10	-0.42	-1.21	-1.89	-2.24	-2.49	-2.67	-2.02	-1.36	-0.75		1.63	4.80	4.51	1.75	0.31
11	1.54	0.75	0.06	-0.28	-0.53	-0.71	-0.07	0.60	1.21	1.63		6.75	6.47	3.70	2.26
12	4.70	3.91	3.23	2.88	2.63	2.45	3.10	3.76	4.38	4.80	6.75		9.63	6.87	5.43
13	4.41	3.63	2.94	2.60	2.35	2.16	2.81	3.48	4.09	4.51	6.47	9.63		6.58	5.14
14	1.65	0.86	0.18	-0.17	-0.42	-0.60	0.05	0.71	1.32	1.75	3.70	6.87	6.58		2.38
15	0.21	-0.58	-1.26	-1.61	-1.86	-2.04	-1.39	-0.73	-0.12	0.31	2.26	5.43	5.14	2.38	

An attempt was made to classify the observations into their corresponding states by using a single PS parameter. This resulted into 51.78% total misclassification error while using Cycle PS parameter, 65.56% total error while using Offset PS parameter and 64.44% total error while using Split PS parameter. This indicates that using only a single PS parameter to classify observations might not be sufficient and might not provide a good classification accuracy.

Three dimensional approach for classifying observations

Using a three dimensional approach for classifying the observations into their corresponding states can yield to a good classification accuracy, as the additional two dimensions can provide additional knowledge about the exactness of the state. For each observation in the data set, we now have three PS parameters (Cycle, Offset and Split) computed and a state label. This data was now used to perform a classification of the observations into different states. Table *12* represents the discriminant functions obtained for the classification. For a particular observation, the state function that yields the highest value is assigned as the state label for that observation.

Table 12 Discriminant functions with all three PS parameters

State	Constant	Coefficient of Cycle PS parameter	Coefficient of Split PS parameter	Coefficient of Offset PS parameter
1	-1.95628	-2.63667	-0.12287	1.3948
2	-20.1762	-8.44358	-0.1457	3.95989
3	-48.8765	-14.3228	2.39014	6.00264
4	-34.249	-11.4188	4.05093	2.74718
5	-38.3097	-11.8307	4.08771	2.52993
6	-46.8768	-12.1208	7.50792	1.33831
7	-32.8519	-8.59532	7.87973	0.01028
8	-29.5407	-7.33178	8.40406	0.28526
9	-23.3769	-6.80161	7.74396	1.33424
10	-29.8932	-8.94012	8.56422	3.44959
11	-21.8678	0.72274	5.56506	1.39598
12	-85.6875	8.73693	3.96261	2.50769
13	-63.1217	4.97033	4.01224	4.59688
14	-24.0363	-1.36917	6.46263	2.83823
15	-1.15825	-0.16775	1.31	0.65993

Table *13* represents the classification summary obtained from SAS. The total misclassification error based on the three PS parameters was reported to be 23.7%, which is significantly less as compared to the error while using a single PS parameter. It can be observed from Table *13* that ten states have more than 75% of their observations classified correctly (highlighted diagonal elements show correct classification into that state), but certain states show cross classifications. This might be largely due to similarity exhibited in terms of traffic characteristics among these states. Hence, a further investigation would be needed to identify as to which states can be combined together based on their similarity, such that it would enhance the classification accuracy without losing the importance of their existence as a separate state. In order to identify as to which states need to be combined based on their similarities, K-means clustering was performed by using mean values of the data representing each group.

Table 13 Classification summary

From State	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
1	**30** **100.00**	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	30 100.00
2	0 0.00	**30** **100.00**	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	30 100.00
3	0 0.00	2 6.67	**26** **86.67**	2 6.67	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	30 100.00
4	0 0.00	0 0.00	1 3.33	**15** **50.00**	14 46.67	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	30 100.00
5	0 0.00	0 0.00	2 6.67	14 46.67	**11** **36.67**	3 10.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	30 100.00
6	0 0.00	0 0.00	0 0.00	1 3.33	0 0.00	**27** **90.00**	2 6.67	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	30 100.00
7	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	3 10.00	**19** **63.33**	7 23.33	0 0.00	1 3.33	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	30 100.00
8	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	9 30.00	**14** **46.67**	5 16.67	2 6.67	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	30 100.00
9	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	3 10.00	5 16.67	**18** **60.00**	4 13.33	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	30 100.00
10	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	2 6.67	4 13.33	**24** **80.00**	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	30 100.00
11	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	**23** **76.67**	0 0.00	0 0.00	7 23.33	0 0.00	30 100.00
12	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	**26** **86.67**	4 13.33	0 0.00	0 0.00	30 100.00
13	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	3 10.00	**27** **90.00**	0 0.00	0 0.00	30 100.00
14	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	6 20.00	0 0.00	0 0.00	**24** **80.00**	0 0.00	30 100.00
15	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	1 3.33	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	**29** **96.67**	30 100.00
Total	30 6.67	32 7.11	29 6.44	32 7.11	25 5.56	33 7.33	33 7.33	28 6.22	28 6.22	31 6.89	29 6.44	29 6.44	31 6.89	31 6.89	29 6.44	450 100.00

Grouping of states based on k-means clustering

k-means clustering algorithm aims to classify a given dataset into certain number of user specified clusters. Figure *5* represents the mean of observations from each state which will be used in the k-means clustering process to find optimal number of clusters and to find which observations (states) can be grouped together.

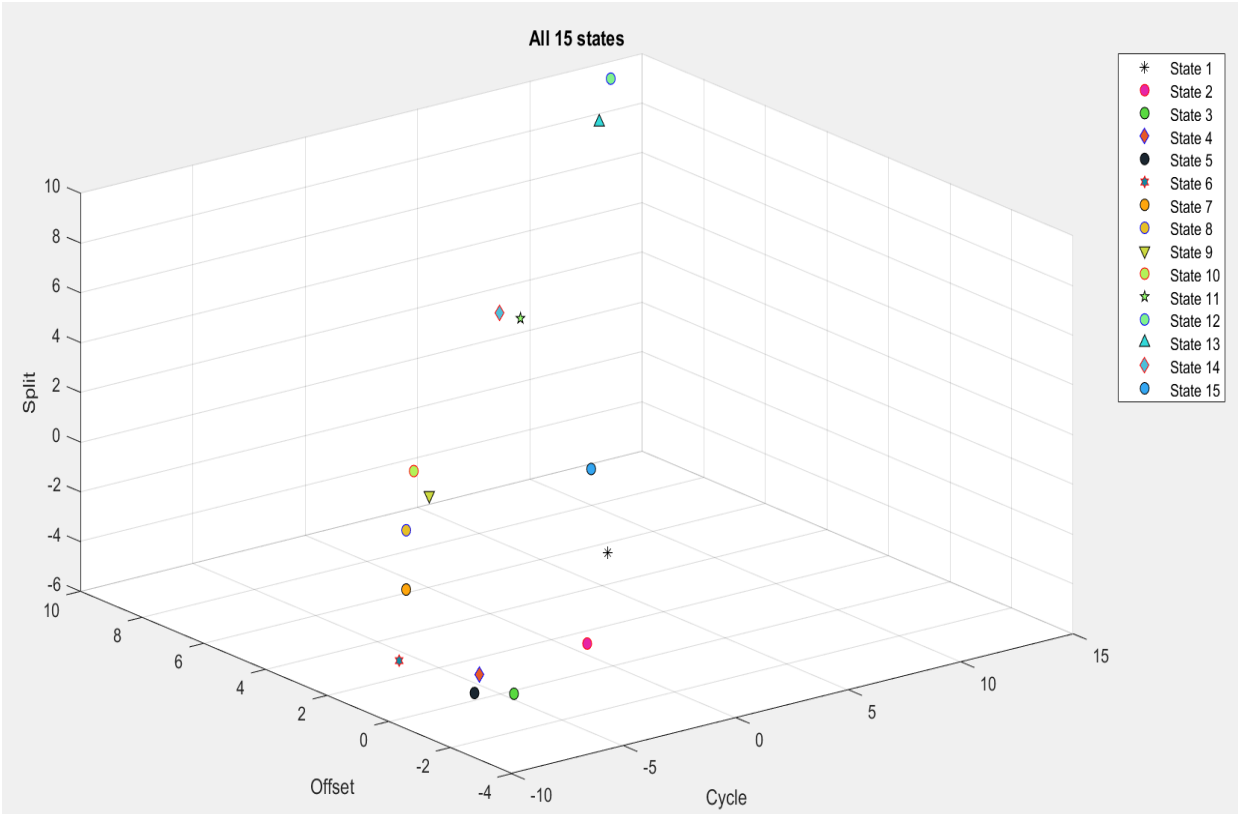

Figure 5 Representation of all 15 states based on three PS parameters

As the number of clusters increase, the total sum of within cluster distance decreases as well. However, it is essential to find the optimal number of clusters. Hence, the k-means algorithm was used to identify the total sum of within cluster distances for cluster sizes ranging from 5 to 14. The lower limit 5 was based on the existence of 5 unique states that had at least 90% of their observations classified correctly (States 1, 2, 6, 13 and 15) which were verified from Table *13* and Figure *5*. Additionally, Silhouette value, which indicates how close each point in one cluster is to the points in the neighboring cluster was calculated. The Silhouette value ranges from -1 to 1, where closeness to 1 indicates that the point is very far away from the neighboring cluster. Hence, mean value of all the clusters would provide us with adequate information as to how well are the clusters separated from each other, while expecting the value to be close to 1. Figure *6* shows a plot of number of clusters (States) against the total sum of within cluster distances and mean Silhouette value.

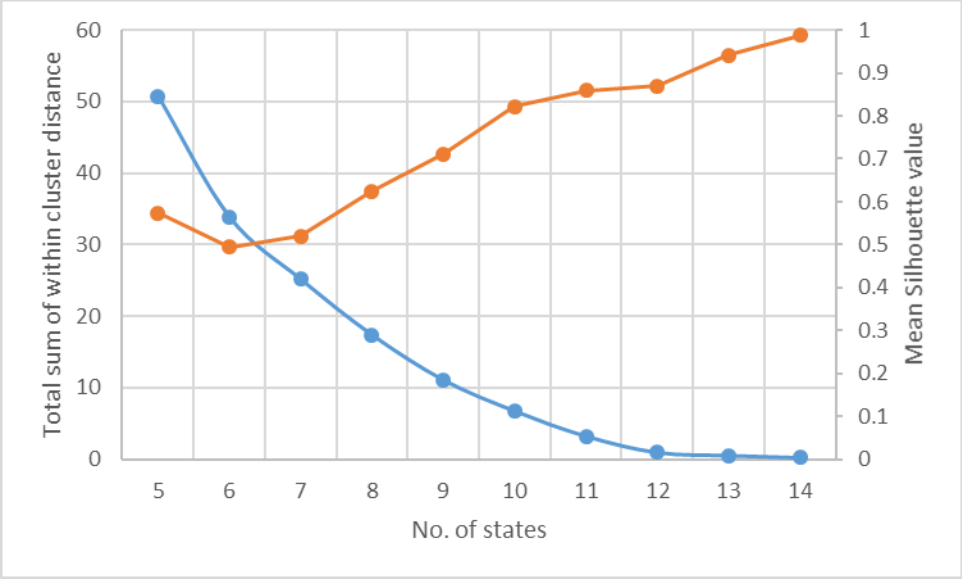

Figure 6 Number of clusters vs Sum of within cluster distance and mean Silhouette value

It can be observed from Figure *6* that as the cluster size increases, the sum of within cluster distance decreases significantly until a point from where the gain is minimal. Similarly, the Silhouette value increases with the number of clusters. It was observed that from cluster size 8, the gain in Silhouette value diminished, however the gain in sum of within cluster distance was significant until cluster size 12. Based on this result, it was decided to analyze the classification accuracy (or misclassification error) for cluster size 8, 9, 10, 11 and 12 by grouping the respective states suggested in each cluster after the k-means process. Table *14* presents a summary of the results for misclassification error against the cluster size.

Table 14 Summary of cluster size vs the classification accuracy

Cluster size (No. of states)	**Total misclassification error (%)**
8	10.42
9	6.28
10	8.44
11	8.48
12	13.33

It can be observed that with 9 clusters (states), the classification accuracy was highest among the considered cluster sizes. This was also supported with a reasonably good Silhouette value and lower total sum of within cluster distances. The 9 states were: 1, 2, (3, 4, 5), 6, 7, (8, 9, 10), (11, 14), (12, 13) and 15. From Table *13* it can be clearly observed that states (3, 4, 5), (8, 9, 10), (11, 14) and (12, 13) show cross classification of observations indicating that they share similar traffic characteristics. It can also be verified from Figure *5* that these states lie very close to each other resulting into them being grouped into a single cluster. Hence, these 9 states are suggested instead of the 15 original states that can produce a classification accuracy of up to 93%.

Analysis at different market penetration rates

As discussed earlier, the queue length data collected were considered as ground truth data (100% penetration rate), and perturbations were introduced into the data at different penetration rates (Table *2*). The discriminant functions developed at 100% penetration rate were used to classify the observations and Figure *7* shows a plot of the total misclassification against the penetration rate.

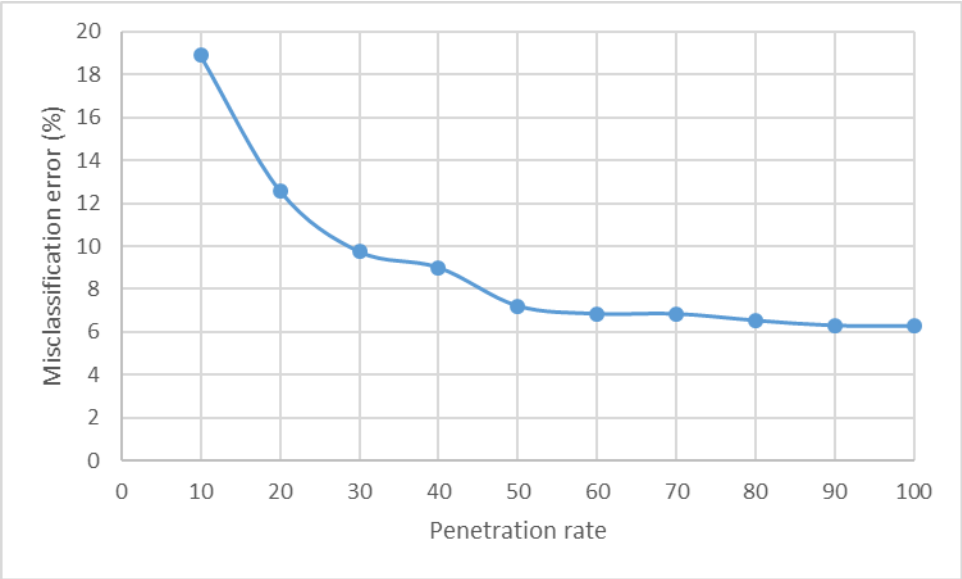

Figure 7 Penetration rate vs Misclassification error

From Figure *7* it can be observed that as the penetration rate is increasing the misclassification error is decreasing which is obvious as the input data is free from errors. However, this result can be used to identify the minimum penetration rate of connected vehicles to be expected in the traffic stream to provide data that can result in accurate state estimation with a certain user defined accuracy. It can be observed from Figure *7* that the gain in classification accuracy becomes marginal beyond 50% penetration rate, indicating that at this rate of market penetration, accurate state estimation can be performed with a reasonable accuracy.

ESTIMATING QUEUE LENGTHS

In the analyses presented above, an indirect method is employed to reflect the impact of market penetration of CVs on the estimated queue lengths (see Table *2*). Alternatively, queue length estimation could be made an integral part of the simulation environment where the market

penetration is a variable. In other words, in each simulation, the market penetration can be varied and a method could be employed to predict the queue lengths in real-time from the data of CVs which serve as probe vehicles. Since this approach will be more complex and computationally demanding it was not employed in this project. However, this section presents a potential method that can be integrated into a microscopic simulation model to predict the queue lengths in real-time in future studies. This method is based on shockwave theory and is developed for under-saturated conditions. For oversaturated conditions, a similar method could be developed as documented in the literature [16].

Methodology

Figure 8 shows a sample shockwave diagram and the critical point Q that needs to be estimated. Point Q represents the maximum extent of the queue for this cycle. The goal is to predict the critical points Q for each cycle. The proposed methodology uses the shockwave theory to determine this unknown point. Essentially, if the speeds of the shockwaves representing the queue growth and dissipation are known, the problem can be solved using basic algebra. These unknown speeds need to be predicted using probe vehicle data. The time-space coordinates when probe vehicles join the back of the queue are denoted by P_j^k in the figure. These coordinates are the main source of data for estimating the shockwave speed and consequently the coordinate of point Q.

To estimate the maximum queue for each cycle k, the probe vehicles (or CVs) joining the back of the queue (if any) are identified. First, a shockwave speed for queue dissipation based on the information obtained from previous cycles in the intersection is measured. Second, by using the probe time stamp and position where it joins the back of the queue (P_j^k in Figure 8), a shockwave speed based on each probe vehicle observation is calculated. Using an exponentially weighted moving average, the average shockwave speed for the probe vehicles arrivals is found. If the arrival of the last probe vehicle is before the end of the current red phase, a shockwave speed based on historical data for non-probe vehicles is also calculated. Using the shockwave speeds for queue formation and dissipation, the coordinates of point Q is determined. Once the coordinates of point Q is known, the queue length is set to Qx, the distance coordinate of point Q. With the estimated Q length, a k nearest neighbor algorithm trained based on historical data is used to detect the number of vehicles stopped behind the stop bar in the observed cycle. Historical data is then updated by adding the current cycle estimations and estimation for next cycle k+1 starts. Here, some notation and general relationships are introduced.

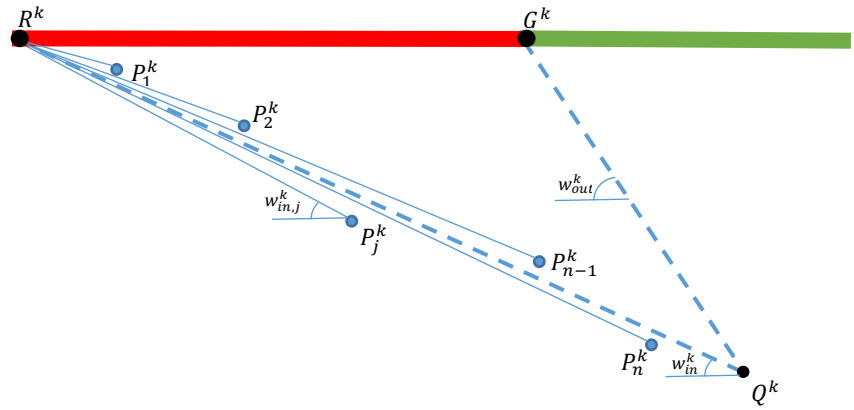

Figure 8 Observed time-space coordinates of sample probe vehicles joining the back of queue and their respective shockwave lines

The shockwave speeds for a probe vehicle at the kth cycle is calculated as:

$$w_{in,j} = \frac{x_j - x_R}{t_j - t_R} \qquad \textit{(Equation 1)}$$

where

x_j, t_j = Space and time coordinates of the probe vehicle when joining the back of the queue

x_R, t_R = Stop-bar location and start time of the red phase, respectively.

Then, the moving average shockwave rate based on the first J probes ($\overline{w}_{in,j}$) is calculated as:

$$\overline{w}_{in,j} = \begin{cases} w_{in,j} & j = 1 \\ \alpha\, w_{in,j} + (1-\alpha)\,\overline{w}_{in,j} & j > 1 \end{cases} \qquad \textit{(Equation 2)}$$

in which the coefficient α represents the weight, a constant smoothing factor between 0 and 1. A higher α discounts older observations faster. α is one of the model parameters which can be optimized using the previously observed data.

After calculating the $\overline{w}_{in,Last}$ for the last probe observation in the kth cycle, the shockwave speed is also smoothed using the shockwave speed of the probe vehicles in k-1th cycle using:

$$\hat{w}_{in}^k = \alpha\, \overline{w}_{in,Last}^k + (1-\alpha)\,\overline{w}_{in,Last}^{k-1} \qquad \textit{(Equation 3)}$$

where $\hat{w}_{in}^k$ is the smoothed inflow shockwave speed. After estimating the shockwave speeds, we can find the coordinates of the interest point Q by intersecting the two shockwave lines.

$$\begin{cases} x_Q^k = x_R^k + \hat{w}_{in}^k (t_Q^k - t_R^k) \\ x_Q^k = x_G^k + \hat{w}_{out}^k (t_Q^k - t_G^k) \end{cases} \qquad \textit{(Equation 4)}$$

Not all cycles contain a probe vehicle. Furthermore, some cycles might only include very few probes. In the case which the last probe vehicle data is observed before the end of red phase, a new shockwave speed, representing the rate at which non-probe vehicles enter the intersection will be calculated to have a more accurate estimation of the queue length. When the last probe vehicle joins the queue before the end of the red phase, the probability that other non-probe vehicles would join the queue before the end of the cycle is greater. To account for the extra non-probe vehicles arriving after the arrival of the last observed probe vehicle, the shockwave speed for the non-probe vehicles in-flow rate is used. Figure 9 illustrates this case. Using simple geometry, the coordinates of point Q in this case can be calculated using:

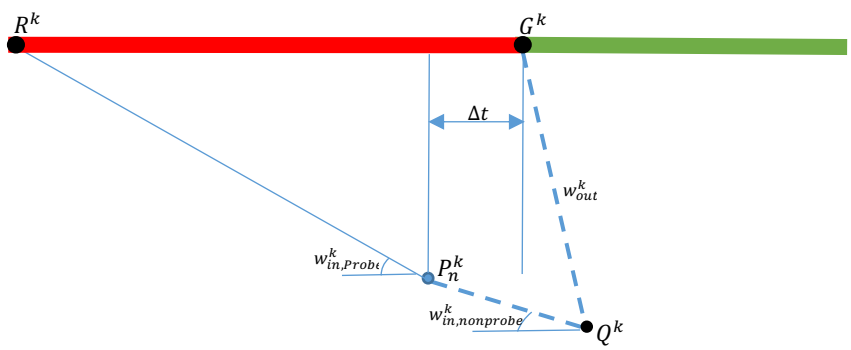

Figure 9- Observed time-space coordinates of probe and non-probe vehicle joining shockwave lines when the last probe arrives before the end of red phase

$$\begin{cases} x_n^k = x_R^k + w_{in,probe}^k (t_n^k - t_R^k) \\ x_Q^k = x_n^k + w_{in,non-probe}^k (t_Q^k - t_n^k) \\ x_Q^k = x_G^k + w_2^k (t_Q^k - t_G^k) \end{cases} \qquad \textit{(Equation 5)}$$

where t_n^k is the time when the last probe joined the back of the queue and the shockwave speed $w_{in,non-probe}^k$ represents the rate at which non-probe vehicles enter the queue and is calculated based on the moving average rate of non-probe vehicles in-flow in previous cycles.

Results

In order to show the application of the formulation given here and to test its performance in estimating queue lengths, a simple network is built in the microscopic simulation software VISSIM. A single one-lane link is created. A traffic signal with 60s of green and 30s of red (cycle length = 90s) is created at 600m location of the link. All vehicles are passenger cars with the desired speed of 55km/h and they enter the network at the rate of the demand profile shown in Figure 10 . Each car is generated via the COM interface of VISSIM and is fed to the simulation to ensure that the input follows the demand profile. The vehicles then follow the car following behavior of Wiedemann 99. Car following characteristics used for this simulation are shown in TABLE 15. Simulation resolution is set to ten times per second (step length = 0.1s). All other values are kept at the default values built within VISSIM.

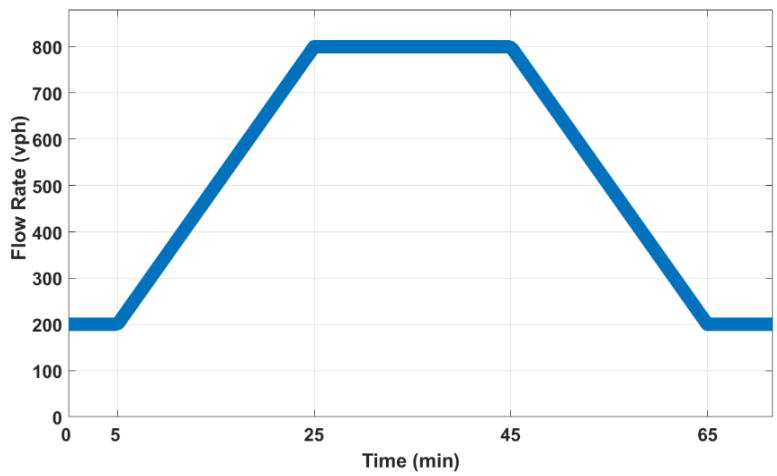

Figure 10-Vehicle Input of the simulation

TABLE 15- Used parameters for Wiedemann 99 car following model

CC0	CC1	CC2	CC3	CC4	CC5	CC6	CC7	CC8	CC9
1.5	0.9	4.0	-8.0	-.035	0.35	11.44	0.25	3.5	1.5

Sensitivity analysis

To evaluate the performance of the formulation developed here, several scenarios are considered. These scenarios are created by varying the available probe vehicle data in the simulation. Each vehicle that enters the network is designated to be probe or not based on a non-fair coin toss (Roulette wheel selection). As the arrivals of the probe vehicles are random, the variability in the arrivals is represented by 50 different simulations for each probe level. In order to have an accuracy measurement of the estimated data, the error in estimation is defined as:

$$Error = \frac{1}{K}\sum_{k=1}^{K}\frac{\left(N^k - N^k_{actual}\right)^2}{N^k_{actual}} \qquad \textit{(Equation 6)}$$

where N^k and N^k_{actual} are the estimated and actual number of vehicles stopped behind in the intersection in the kth cycle, respectively. K is the total number of cycles for estimation, which is 50 cycles in this article.

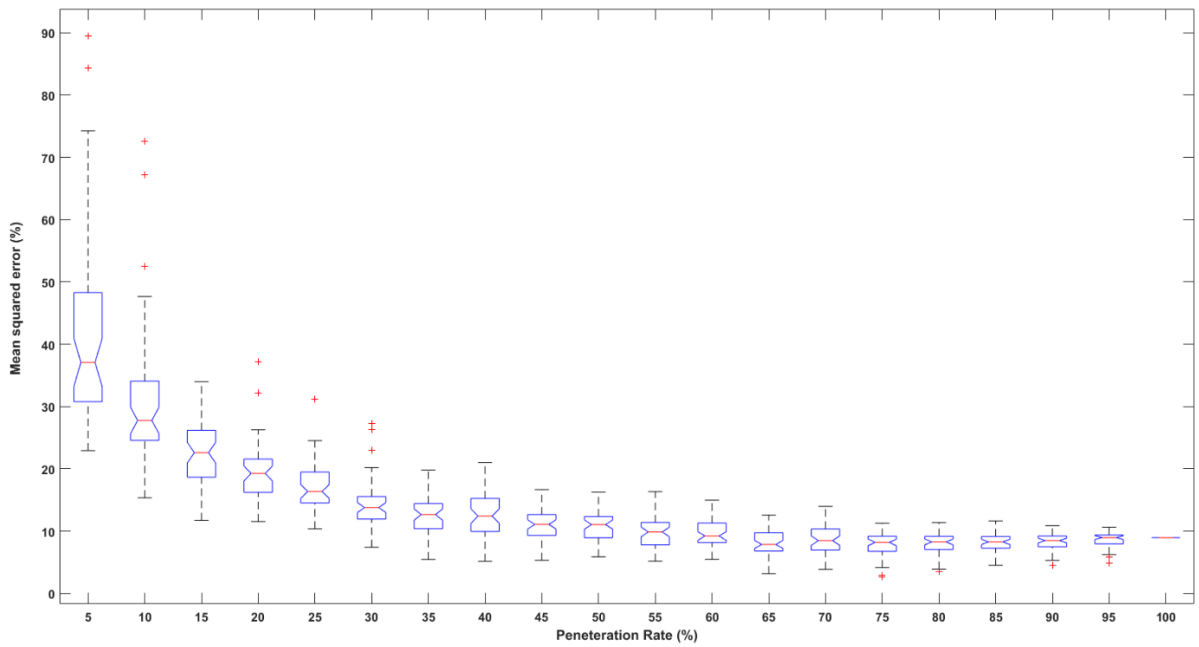

Figure 11- Notched boxplot of sensitivity analysis

Figure 11 represents the notched box plot of the reduction in the error percentage as penetration rate increases. The notches in this diagram represent as the market penetration rate increases, the error in estimation decreases. However, it can be seen that beyond a relatively small penetration rate of 30%, the mean estimation error is not improved significantly. Moreover, the variations of error percentage in low market penetration rates, i.e. 5%, is substantially higher than those of higher penetration rates. In other words, we can be more certain about the mean error percentage as the penetration rate increases.

In order to better understand the reasoning behind the variations in errors for different penetration rates, more in-depth study for each case has been carried out. As it can be seen from Figure 11, the highest variance is for 5% penetration rate where we can have error percentages as high as 90% or as low as 23%. Figure 12 shows the comparison between estimated queue lengths obtained from probe vehicle data (blue lines) and ground truth (diamond dots). As it can be seen in this figure, with an increase in the moving average of input vehicles starting at cycle number 8, in the best case scenario is able to catch up with this increase and can predict the increment in queue length amount accordingly. Although there are observable lags in adaptation of the model's prediction based on the observed change in vehicle input average which is represented by the flat lines in the figure, the model is performing relatively well.

In the worst case scenario, however, the model mistakenly predicts a low queue length at cycle number 15 and after 3 whole cycles starts to see the increase in vehicle input. The worst case

persistently predicts a greater queue length albeit the decrease in demand after the 30th cycle. This is because no new probe vehicle is stopped behind the signal to update the model's estimation parameters.

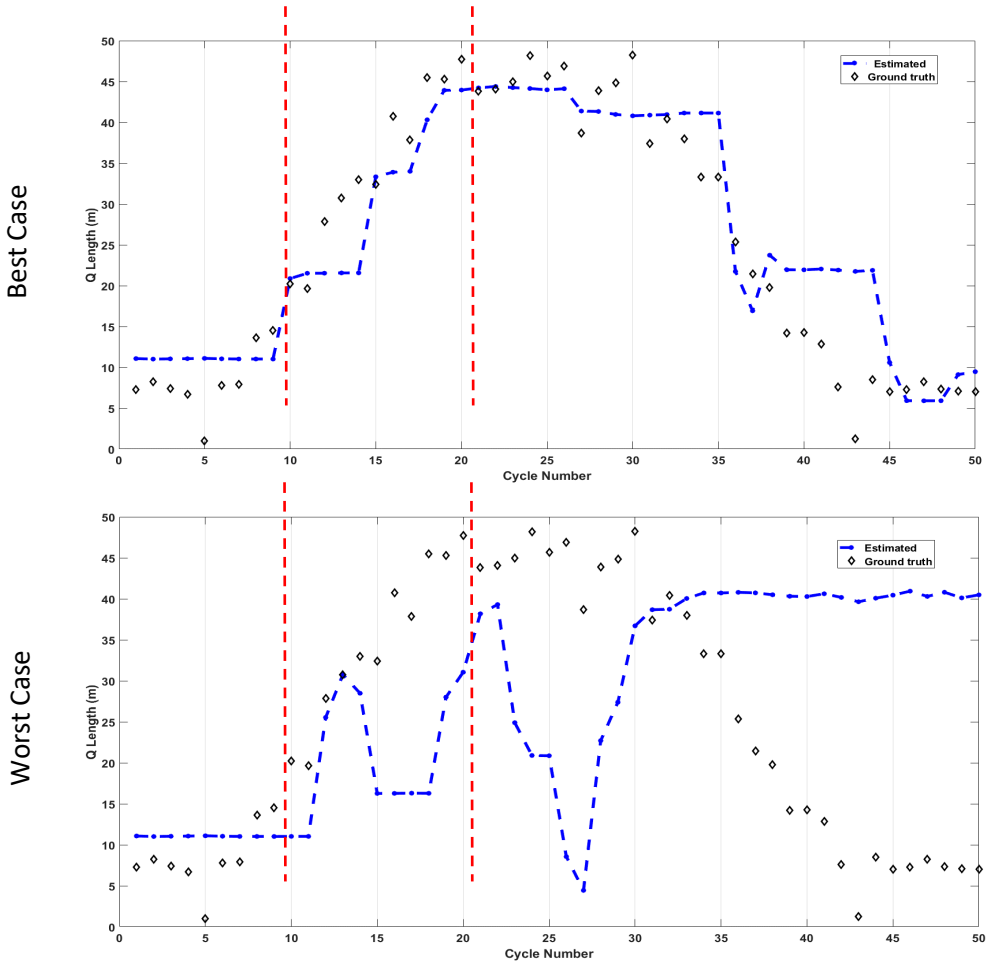

Figure 12 - Comparison between the best and worst case queue length estimations at market penetration rate of 5% for 50 cycles of simulation

In the next step, we can look at the trajectories of the probe vehicles joining the back of the queue for cycles 9 through 20 (the area between the two dashed red lines in Figure 12), where the vehicle input is increasing. Figure 13 depicts the probe vehicle trajectories (red lines), queue length predictions (black lines), and ground truth queue lengths (green lines) for the best and worst case estimation scenarios at the market penetration rate of 5%. First, in the worst case scenario, more than half the vehicles do not stop behind the signal. At the same market penetration rate, a lower number of probe vehicles stopped behind the signal will provide less information to the model for estimation purposes. Second, in the worst case scenario there is a condensed density of probe vehicle arrivals in the short period between 12th and 14th cycle and no probe vehicle data obtained by the signal from 14th through 19th cycle. This non-homogenous distribution of probe vehicles leads to a greater number of blank cycles (cycles without any new information obtained from probe vehicles). Whereas in the best case scenario, a relatively uniform distribution of probe vehicle arrival is observed. Third, the stopped probe vehicles in the best case scenario have joined the back

of the queue at a later point in the cycle, providing more accurate information about the queue length. For example, the probe vehicles joining the queue in cycles 10, 15, and 18 are essentially giving the exact value of queue length in the best case, whereas the probe vehicle in cycles 12, 14 or 19 of the worst case have joined in the middle of the queue.

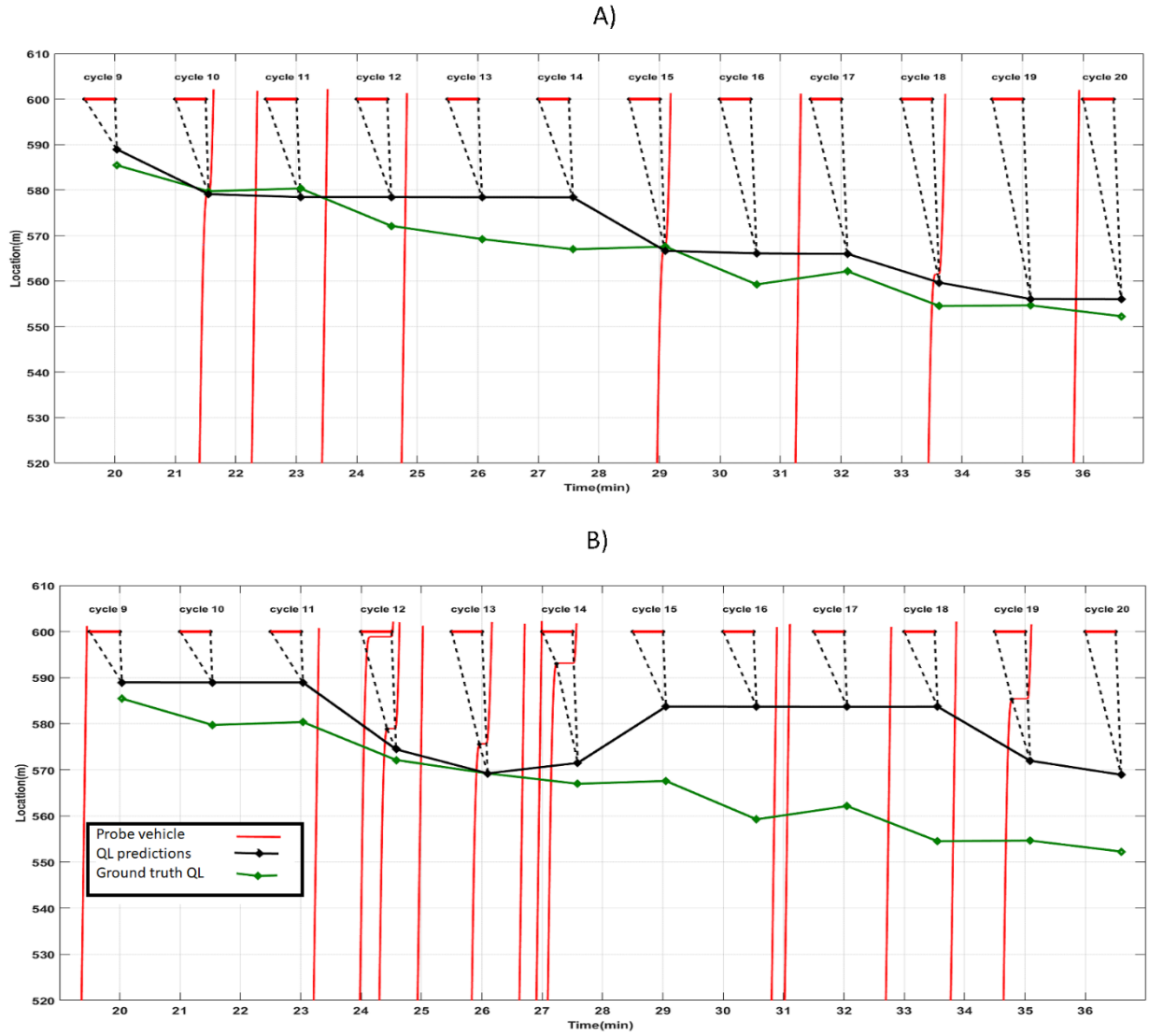

Figure 13 – Comparison between A) the best and B) the worst case vehicle trajectories for prediction at market penetration rate of 5%

SUMMARY AND CONCLUSIONS

In the current study, a framework to obtain counts, occupancy and queue lengths from the VISSIM network was developed. The study focused on estimating three PS parameters. This was done by selecting a particular set of detectors and queue counters for each PS parameter. The

best set of timing plans were selected based on total queue delay and these were used to obtain data for all the traffic states. Canonical discriminant analysis was performed to obtain weights associated with each variable for each PS parameter based on which the three PS parameters were obtained. These were then used to obtain discriminant functions to classify the observations into different states.

Following a reasonable classification accuracy, k-means clustering approach was used to reduce the number of states by clustering similar traffic states together. Based on this analysis, 9 states were suggested instead of the original 15 states for which a 93% classification accuracy was obtained at 100% penetration rate. The developed functions were then used to perform classifications for data at different penetration rates and the corresponding misclassification errors rates were reported. It was observed that the gain in classification accuracy diminished at 50% penetration rate. Overall, from this study it was demonstrated that queue length data can be a valuable source of information for traffic state estimation for implementation in TRPS framework. For future studies, the queue length could be estimated in real-time from the data provided by connected vehicles in the traffic stream, and these estimates could be used directly for system state estimation to support TRPS implementations. The report provided a potential queue length estimation method based on shockwave speeds and showed how the accuracy varies with market penetration rate of connected vehicles. Other methods for queue length estimation could also be considered in future studies.

REFERENCES

1. Box, S., et al., *Urban traffic state estimation for signal control using mixed data sources and the extended Kalman filter.* 2013.
2. Liu, H.X. and X. Di, *Development of Algorithms for Travel Time-Based Traffic Signal Timing, Phase I–A Hybrid Extended Kalman Filtering Approach for Traffic Density Estimation along Signalized Arterials.* 2010.
3. Srinivasan, D., M.C. Choy, and R.L. Cheu, *Neural networks for real-time traffic signal control.* IEEE Transactions on Intelligent Transportation Systems, 2006. **7**(3): p. 261-272.
4. Khan, S.M., K.C. Dey, and M. Chowdhury, *Real-Time Traffic State Estimation With Connected Vehicles.* IEEE Transactions on Intelligent Transportation Systems, 2017.
5. Mannini, L., et al., *On the short-term prediction of traffic state: an application on urban freeways in Rome.* Transportation Research Procedia, 2015. **10**: p. 176-185.
6. Abbas, M.M., et al., *Methodology for determination of optimal traffic responsive plan selection control parameters.* Research Report, 2003.
7. Abbas, M. and S. Abdelaziz, *Evaluation of Traffic Responsive Control on the Reston Parkway Arterial Network*. 2009, Virginia Center for Transportation Innovation and Research.
8. Abbas, M.M. and A. Sharma, *Optimization of time of day plan Scheduling using a multi-objective Evolutionary algorithm.* 2005.
9. Abbas, M. and A. Sharma, *Configuration of traffic-responsive plan selection system parameters and thresholds: Robust bayesian approach.* Transportation Research Record: Journal of the Transportation Research Board, 2004(1867): p. 233-242.

10. Abbas, M., et al., *Configuration methodology for traffic-responsive plan selection: A global perspective.* Transportation Research Record: Journal of the Transportation Research Board, 2005(1925): p. 195-204.
11. Sharma, A., *Determination of traffic responsive plan selection factors and thresholds using artificial neural networks*. 2004, Texas A&M University.
12. Li, J.-Q., et al., *Estimating queue length under connected vehicle technology: Using probe vehicle, loop detector, and fused data.* Transportation Research Record: Journal of the Transportation Research Board, 2013(2356): p. 17-22.
13. Badillo, B.E., et al. *Queue length estimation using conventional vehicle detector and probe vehicle data*. in *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*. 2012. IEEE.
14. Tiaprasert, K., et al., *Queue length estimation using connected vehicle technology for adaptive signal control.* IEEE Transactions on Intelligent Transportation Systems, 2015. **16**(4): p. 2129-2140.
15. Argote, J., et al. *Estimation of measures of effectiveness based on Connected Vehicle data*. in *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*. 2011. IEEE.
16. Cetin, M., *Estimating Queue Dynamics at Signalized Intersections from Probe Vehicle Data:Methodology Based on Kinematic Wave Model.* Transportation Research Record, 2012. **2315**(1): p. 164-172.

APPENDIX I: Timing Plans

Table 16 Phase times (splits) for controller 1010

Plan	1	2	3	4	5	6	7	8
1		45		30	13	32		
2		45		30	13	32		
3		60		30	13	47		
4		60		30	13	47		
5		60		30	13	47		
6		69		31	13	56		
7		70		30	13	57		
8		70		30	13	57		
9		111		39	15	96		
10		118		32	16	102		

Table 17 Phase times (splits) for controller 1011

Plan	1	2	3	4	5	6	7	8
1	12	32		31	17	27		31
2	12	32		31	15	29		31
3	12	45		33	20	37		33
4	12	42		36	17	37		36
5	12	40		38	18	34		38
6	12	51		37	22	41		37
7	12	47		41	20	39		41
8	12	46		42	20	38		42
9	12	81		57	34	59		57
10	12	73		65	28	57		65

Table 18 Phase times (splits) for controller 1012

Plan	1	2	3	4	5	6	7	8
1	12	31	12	20	12	31	12	20
2	19	27	13	16	12	34	12	17
3	14	39	14	23	12	41	12	25
4	18	37	13	22	12	43	12	23
5	21	36	14	19	15	42	12	21
6	16	44	14	26	12	48	12	28
7	21	41	14	24	12	50	12	26
8	23	41	16	20	16	48	12	24
9	25	67	20	38	16	76	12	46
10	36	63	25	26	23	76	13	38

Table 19 Phase times (splits) for controller 1013

Plan	1	2	3	4	5	6	7	8
1	12	32		19	12	32		12
2	12	39		12	12	39		12
3	12	38		28	12	38		12
4	12	36		25	12	36		17
5	12	48		18	12	48		12
6	12	42		34	12	42		12
7	12	40		30	12	40		18
8	12	53		21	12	53		14
9	12	55		68	12	55		15
10	12	77		43	12	77		18

Table 20 Phase times (splits) for controller 1014

Plan	1	2	3	4	5	6	7	8
1	17	30	12	16	25	22		28
2	17	26	12	20	14	29		32
3	19	39	12	20	31	27		32
4	21	37	12	20	26	32		32
5	18	34	12	26	17	35		38
6	20	45	12	23	35	30		35
7	21	45	12	22	30	36		34
8	18	39	12	31	18	39		43
9	32	66	15	37	54	44		52
10	27	62	15	46	27	62		61