Summer 2015

Characterizing Queue Dynamics at Signalized Intersections From Probe Vehicle Data

Semuel Yacob Recky Rompis
Old Dominion University

CHARACTERIZING QUEUE DYNAMICS AT SIGNALIZED INTERSECTIONS FROM PROBE VEHICLE DATA

by

Semuel Yacob Recky Rompis
B.E. August 2000, Sam Ratulangi University, Indonesia
M.Eng. September 2003, Sam Ratulangi University, Indonesia
M.Eng. August 2009, University of South Australia, Australia

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

CIVIL AND ENVIRONMENTAL ENGINEERING

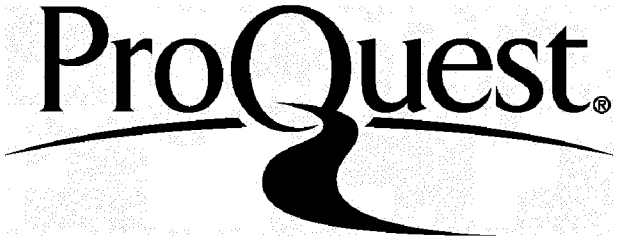
OLD DOMINION UNIVERSITY
August 2015

Approved by:

Mecit Cetin (Director)

ManWo Ng (Member)

Rajesh Paleti (Member)



ABSTRACT

CHARACTERIZING QUEUE DYNAMICS AT SIGNALIZED INTERSECTIONS FROM PROBE VEHICLE DATA

Semuel Yacob Recky Rompis
Old Dominion University, 2015
Director: Dr. Mecit Cetin

Probe vehicles instrumented with location-tracking technologies have become increasingly popular for collecting traffic flow data. While probe vehicle data have been used for estimating speeds and travel times, there has been limited research on predicting queuing dynamics from such data. In this research, a methodology is developed for identifying the travel lanes of the GPS-instrumented vehicles when they are standing in a queue at signalized intersections with multilane approaches. In particular, the proposed methodology exploits the unequal queue lengths across the lanes to infer the specific lanes the probe vehicles occupy. Various supervised and unsupervised clustering methods were developed and tested on data generated from a microsimulation model. The generated data included probe vehicle positions and shockwave speeds predicated on their trajectories. Among the tested methods, a Bayesian approach that employs probability density functions estimated by bivariate statistical mixture models was found to be effective in identifying the lanes. The results from lane identification were then used to predict queue lengths for each travel lane. Subsequently, the trajectories for non-probe vehicles within the queue were predicted. As a potential application, fuel consumption for all vehicles in the queue is estimated and evaluated for accuracy. The accuracies of the models for lane identification, queue length prediction, and fuel consumption estimation were evaluated at varying levels of demand and probe-vehicle

market penetrations. In general, as the market penetration increases, the accuracy improves. For example, when the market penetration rate is about 40%, the queue length estimation accuracy reaches 90%. The dissertation includes various numerical experiments and the performance of the models under numerous scenarios.

This dissertation is dedicated to my wife Lidya and my son Jonathan
and for the honor and glory of Christ the Lord.

ACKNOWLEDGEMENTS

My thanksgiving and gratitude go to Jesus Christ the Lord as I believe He is the one who has strengthened me throughout all the process of this PhD study.

I would like to express my deepest appreciation to my academic advisor, Dr. Mecit Cetin for the academic and financial support, and endorsement all the way through this PhD study, whom without his assistance, support and guidance, the writing and completion of this Dissertation would not have been possible.

I would like to thanks my committee members, Dr. ManWo Ng and Dr. Rajesh Paleti for their guidance. Thanks also to Dr. Asad Khattak for sharing the knowledge and skill while I studied in the Old Dominion University.

Thanks are extended to the Indonesian government and the USA government through the Fulbright Scholarship that has given me the chance to study in Norfolk, Virginia. Thanks are also extended to the Transportation Research Institute, Old Dominion University for providing such a good knowledge to develop skills in Transportation Engineering field.

Special thanks to all the friends in Norfolk; Indonesian friends, American friends, International friends and Fulbrighter friends for their support that has been given during my studies. To my extended family in Manado, Indonesia, I would like to thank for the support and encouragement.

Finally, I would like to express my appreciation and also gratefulness to my wife Lidya and my son Jonathan. This Dissertation is dedicated for both of them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

The high cost of transportation infrastructure construction has triggered a need to maximize the efficiency of the existing transportation system notably in big cities. The goal is to be able to minimize, if not solve, transportation problems such as traffic jams, delays, and safety. Further, it is also required to have the capability to make the transportation system more environmentally friendly. This expectation has made reliable transportation data become more essential; however, using the existing transportation instrument, it is very costly to obtain the data, especially at the network level, which has always been a problem. Hence, researchers and traffic engineers have tried to find alternative devices that are able to provide reliable, but also affordable, transportation data.

Advances in information and technology have made it possible to collect comprehensive data from instrumented vehicles, also known as probe vehicles. The data collected by the probe vehicle system could be very detailed because of the ability to collect second-by-second traffic parameter information.

The Federal Highway Administration (FHWA) [1] categorized five types of ITS probe vehicle data collection: Signpost-Based Automated Vehicle Location (AVL), Automatic Vehicle Identification (AVI), Ground-Based Radio Navigation, Cellular Geolocation, and Global Positioning System (GPS). The advancement of the satellite technology for communication systems has made the use of probe vehicle systems with GPS instruments preferable among all

these types. The probe vehicles are capable of reporting their timely trajectories and speed in the link or network which are required to accurately estimate essential traffic parameters, such as queue length and even fuel consumption and emission. consumed and generated by each individual vehicle. Furthermore, wireless communication advancement has made it possible to have more and more mobile instruments such as smartphones, GPS navigators, etc. perform as the probe instrument.

Although this system is very promising, there are two main issues that need to be addressed. The classic one is the instrument vehicle market penetration rate. In the early years of probe vehicle utilization, the market penetration rate has become an inevitable issue as it is very expensive to install instruments for most, if not all, of the vehicles in the network. However, as explained before, the improvement in information and technology has made it possible to have a higher probe vehicle market penetration rate. Still the problem is not solved. The reason behind this is the privacy issue; not everybody is willing to share their vehicle's time and space information and thus it is still hard to have a high market penetration rate.

The challenge for the traffic engineer is how to conduct reliable traffic parameter estimation using a low/acceptable probe market penetration rate. Another substantial problem with this system is the instrument level of accuracy. While it is capable to report vehicle position at every second, it also has accuracy error of about 3 to 15 meters. The significance of this issue really depends on the objective of using the probe data. It might not be a big problem for traffic macroscopic modeling, but when it comes to a more detailed level, this error becomes a major problem. For example, when conducting research on lane identification, this GPS instrument issue will need to be considered seriously.

Most probe vehicle research has been carried out to predict the travel time, for example [2-4]. Beside travel time prediction, a probe vehicle has been used for other traffic parameter prediction such as estimating queue length [5] and even incident detection [6, 7]. Recently, although only in limited numbers, researchers have conducted studies for utilizing a probe vehicle system to estimate fuel consumption and emission as demonstrated in [8]. This queue length, fuel consumption and CO_2 estimation research was conducted by simplifying the lane problem which is related to the GPS error problem. The assumption is that either the estimation is for a single lane, or the lane is known or the queue is equal for multiple lanes, which obviously are not always true in real life.

This research was conducted to estimate the queue length and fuel consumption by taking into account the market penetration rate and the GPS instrument issue. To be more specific, this research is about lane-based queue length estimation and lane-based fuel consumption and CO_2 estimation using the probe vehicle data.

1.2 Problem Definition

There have been few studies reported about queue length and fuel consumption estimation using probe vehicles, particularly the GPS instrumented vehicle. As explained before, those previous studies were conducted while overlooking the GPS error problem.

Basically, this study is an attempt to optimize the probe vehicle system in predicting queue length and fuel consumption in an urban network, which in this case, is at a signalized intersection. This was done by identifying the lane of the probe vehicle to properly carry out the estimation. The challenge in this study is to find an appropriate technique in detecting a probe

vehicles' lane, by considering the instrument error, using the traffic flow theory and a statistical method.

Furthermore, as mentioned earlier, an essential issue in probe vehicle implementation is the market penetration rate of probe vehicles which has also been addressed by [9, 10]. Logically, the higher the rate, the more accurate the result expected. However, it is not an easy task to apply a high probe vehicle penetration rate in real life. Therefore this issue, for the case of probe vehicle usage in queue length and the fuel consumption estimation in particular, will be investigated in this study.

1.3 Research Questions

This study was conducted to answer the following research questions:

- What is the best methodology for real-time probe vehicles' lane identification at multiple-lane signalized intersections in the case of unequal queues?
- Given the previous question answered, what is the applicable method to estimate the queue length and fuel consumption properly?
- At what rate can the probe market penetration be considered adequate to carry out 1) the probe vehicles' lane identification, 2) the queue length estimation and 3) the fuel consumption estimation?

1.4 Research Objective and Scope of Study

The essential objective of this research is to find an appropriate way in real-time identification of the probe vehicle lane so that data can be used to optimize the queue length and fuel consumption estimation in an urban network. The study is expected to give a good idea of the acceptable market penetration rate that is adequate to carry out the estimation properly.

Although this study is an attempt to find an appropriate method for improving queue length and fuel consumption estimation, it is not within the scope to create a new fuel consumption and emission model. Instead, this study focuses on finding a methodology for optimizing the use of probe vehicles in estimating traffic parameters and fuel consumption.

1.5 Research Contribution

In this dissertation, various models have been developed and analyses were performed to investigate how probe vehicle data could be used to accurately, effectively, and efficiently estimate queue length and fuel consumption at signalized intersections with multiple lanes. The following are the key contributions of this research:

1. This research, to the best of the author's knowledge, is the first attempt to estimate traffic parameters from probe vehicle data while considering its travel lane at signalized intersections. In previous studies, lane identification is either ignored or models are developed for single-lane roads.
2. New models, utilizing traffic flow theory and Bayesian inference, were developed to estimate the lane of the probe vehicle and queue lengths for each individual lane.
3. Based on empirical data from microsimulation, this research evaluated the best set of variables to be extracted from probe vehicle trajectories for lane estimation at signalized intersections with unequal queue lengths.
4. This research is one of a very limited number of studies on estimating fuel consumption from probe vehicle trajectories.
5. The dissertation documents a comprehensive set of simulation experiments and analyses for evaluating effectiveness of different methods for lane prediction, queue length, and fuel consumption estimation.

1.6 Dissertation Proposal Layout

This dissertation consists of six chapters. Chapter 2 contains the literature review; in this chapter several theory and related studies about probe vehicle utilization, queue length estimation, and vehicle fuel consumption emission estimation were reviewed. In Chapter 3, the methodology for lane identification is proposed which includes case studies for several queuing scenarios. The implementation of the lane identification for the queue length estimation is provided in Chapter 4. In Chapter 5 the application of lane identification for fuel consumption is covered. Finally, the conclusion of the previous chapters and recommendations for future research are given in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

2.1 Probe Vehicles

Probe vehicles are becoming an increasingly important topic in transportation systems. The system itself has been studied extensively both in real life and virtually (using transportation simulation software). Research has been conducted to explore the utilization of probe vehicle systems for travel time estimation [2, 4, 11], incident detection [6, 7], lane change maneuver detection [12], traffic states evaluation [13, 14], queue length estimation [5, 15-17], route choice behavior and origin-destination estimation [18], and even for pavement roughness estimation [19, 20].

There are several kinds of probe vehicle systems based on their data transmission approaches. As stated in [13], there are three ways by which probe vehicles can transmit traffic information: 1) space-based, where the traffic information is transmitted to roadside devices as the probe vehicles pass observation points. These kind of data can be retrieved from a beacon-based probe system, an electronic toll tag system or an automatic license plate recognition system; 2) time-based, where the traffic information is reported at every specific time instant wherever the probe vehicles are. These types of data can be obtained from a GPS-based system or a beacon-based system; and 3) event-based, where traffic information is reported as a particular event (such as a traffic incident) occurs.

2.2 The Advantages and Disadvantage of Using Probe Vehicle Systems

The FHWA in the Travel Time Data Collection Handbook [1] describes the common advantages and disadvantages of using probe vehicles systems for travel time data collection, as shown in Table 1. Since the data needed for conducting this study are the same as for travel time data collection, the advantages and disadvantages mentioned in Table 1 also apply to this study.

Table 1. The profit and consequences of using probe vehicle systems

Advantages	Disadvantages
Low cost per unit data	High implementation cost
Continuous data collection	Fixed infrastructure constraints
Automated data collection	Requires skilled software designers
Data are in electronic format	Privacy issues
No disruption of traffic	Not recommended for small data collection efforts

The probe vehicle system has a high initial cost to install the necessary equipment and train personnel to operate it; however, once it has been done, the data could be collected easily at a low cost. Another disadvantage of this system is that it is very costly to adjust the supporting infrastructure (such as the receiving antennas) in the size and system coverage area; thus, before implementation, a proper study will need to take place. This system also will need skilled personnel to operate, gather, and analyze the data. Another issue about probe vehicles is drivers' privacy. As this system involves tracking the location of drivers, motorists may think that their

travel behavior is being monitored. Since this system requires an initial high cost, it is most cost-effective to be implemented within a large study area.

2.3 Probe Vehicle Data Collection System

As has been mentioned in the previous chapter, according to FHWA [1], there are five types of ITS probe vehicle data collection techniques:

Table 2. Probe vehicle data collection systems

Technique	Costs				Data Accuracy	Constraints	Driver Recruitment
	Initial costs	Installation	Data Collection	Data Reduction			
AVL	High	High	Low	High	Low	No. of signpost sites, transit routes, and probes	None; uses transit vehicle
AVI	High	High	Low	Low	High	No. of antenna sites and tag distribution	Required, but can use toll patrons
Ground-Based Radio Navigation	Low	Low	Low	Low	Moderate	No. of probes and size of service area	Required
Cellular Geolocation	High	High	Low	Moderate	Low	No. of cell users and cell towers	None; uses current cellular users
GPS	Low	Low	Low	Moderate	High	No. of probes	Required, but can also use currently instrumented vehicles

Source :[1]

- Signpost-Based Automatic Vehicle Location (AVL), where the probe vehicles communicate with transmitters installed on an existing signpost structure;
- Automatic Vehicle Identification (AVI), where electronic tags are installed in the probe vehicle so that the probe vehicle can communicate with roadside transceivers for identification and for collecting travel time between transceivers;
- Ground-Based Radio Navigation, where the data are collected through probe vehicles and radio tower communication;
- Cellular Geo-location, where cellular telephone call transmissions are tracked discreetly; and
- Global Positioning System (GPS), where probe vehicles are equipped with the GPS transmitters and receivers. This system works with the assistance of satellite service.

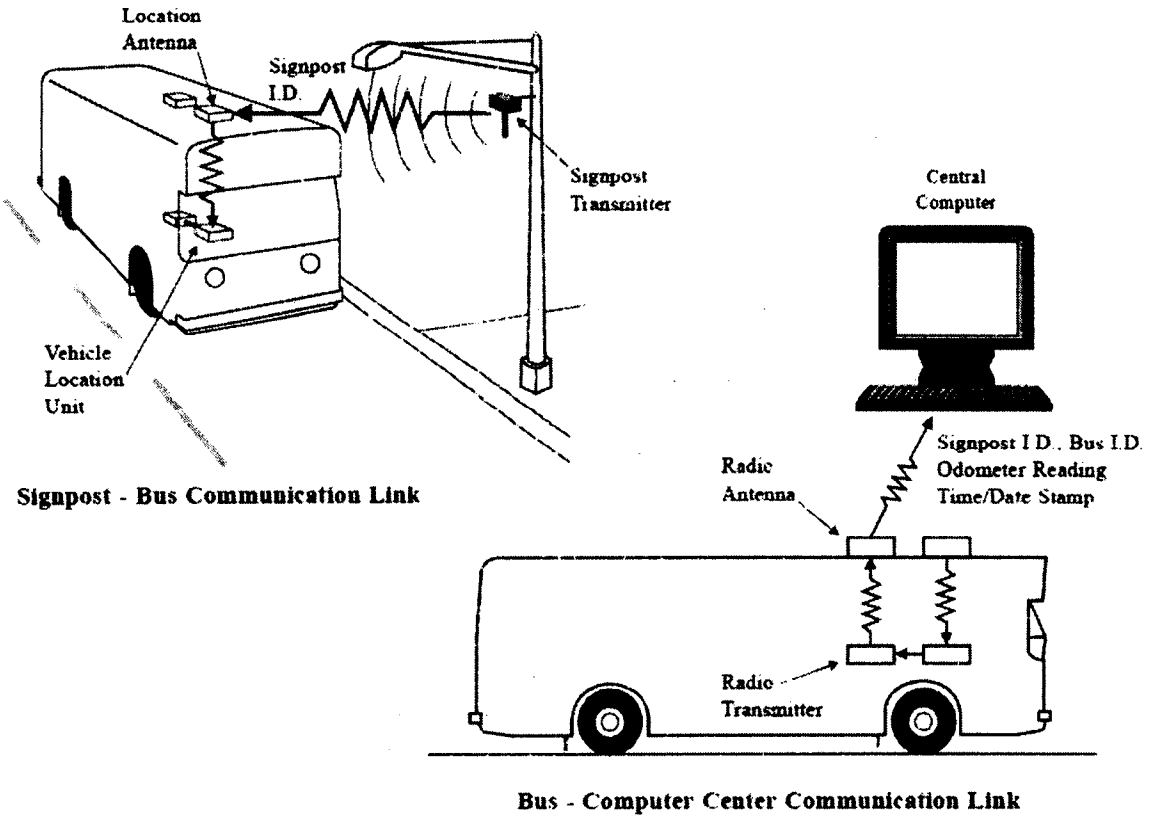
Table 2 demonstrates the comparison between those systems.

2.3.1 Signpost-Based Automatic Vehicle Location (AVL)

There are seven main components to collect and store travel time data in this system:

1. Infrastructure of electronic transmitters,
2. In-vehicle receiver,
3. In-vehicle odometer sensor,
4. In-vehicle locating unit, or data microprocessor,
5. In-vehicle radio transmitter,
6. Central control radio receiver, and
7. Central control facility

The communication process of signpost-based AVL is shown in Figure 1. This system has the following advantages: 1) simple infrastructure; 2) some types of this system are capable of collecting vehicle performance data such as fuel consumption, oil pressure or cooling temperature; 3) some types of this system would be able to collect passenger count; these data will be useful for developing the origin-destination matrix in trip generation studies.

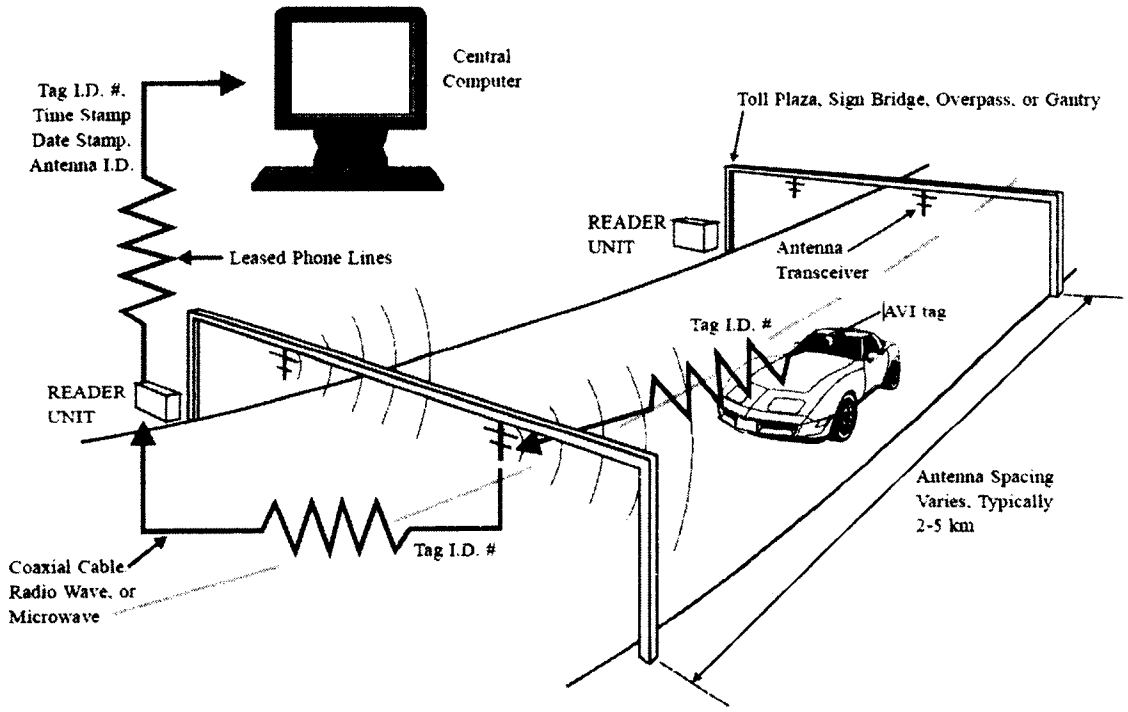


Source :[1]

Figure 1. Signpost-Based AVL system

This system suffers from these disadvantages: 1) since it was originally designed to monitor transit fleet operations, travel time data may not be acceptable to be used in regular traffic studies; 2) with the same reason as the previous, this system covers only data that are limited to

roadways traveled by transit vehicles; 3) routine calibration will be needed in order to have accurate data; 4) advances in the field of satellite technology have caused transit agencies to upgrade their systems to GPS technology that is more accurate and robust; and 5) the data produced from this system require extensive editing and quality control, sometimes even spot-editing (particularly from poorly placed or malfunctioning signpost transmitters).



Source : [1]

Figure 2. AVI System

2.3.2 Automatic Vehicle Identification (AVI)

There are four components (illustrated in Figure 2) that are needed in order to make the automatic vehicle identification (AVI) system work:

1. ITS probe vehicle with electronic transponders,
2. Roadside antenna to detect signal transmitted from electronic transponders,
3. Roadside readers to collect the data, and
4. Central computer for collecting and processing all data.

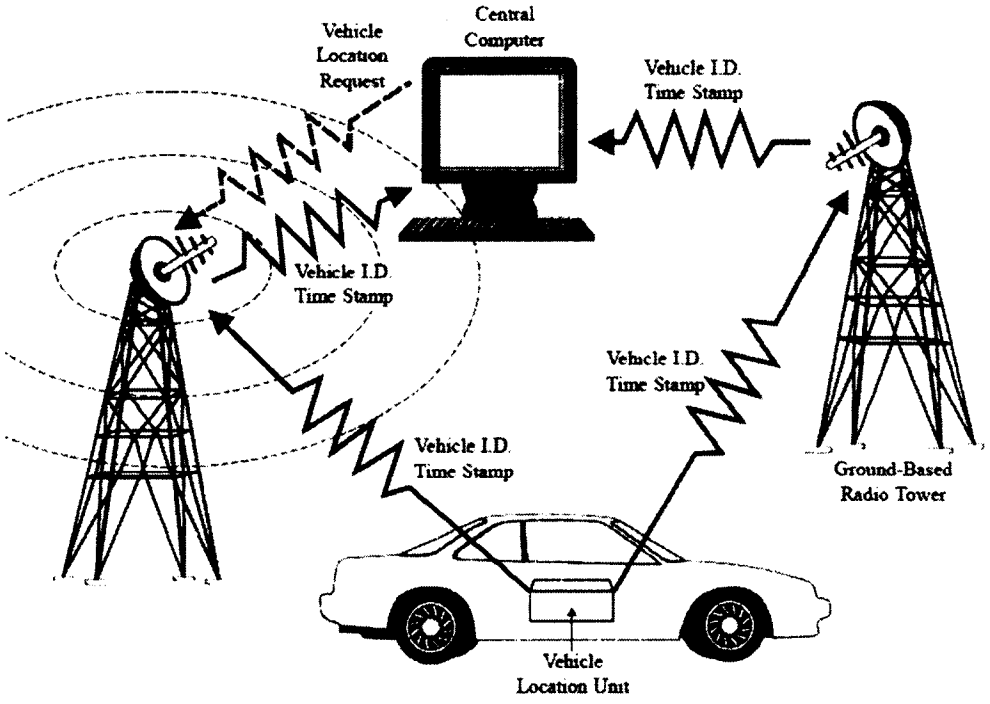
Compared to the other ITS probe vehicle systems, the AVI systems benefit from these advantages: 1) AVI systems have the capability to collect data continuously for an entire 24-hour period; 2) as this system is completely automated, very few human resources are needed to make this system work; therefore 3) this system has minimum risk of injury and human error potential in the data collection process; 4) this system has been proven to have better accuracy in collecting the data and also immune to interference from cellular telephones, commercial radio signal and electric generators; 5) this system is able to collect data from specific lanes; and 6) this system is also appropriate to collect vast amounts of data, for example, data collection for over a year in all types of environmental situations. The disadvantages of this system are 1) it depends heavily on infrastructure, i.e. the system can only collect data from roads within the coverage area of AVI infrastructure (such as antennae of ETC booths); 2) data can only be collected from vehicles with tags in use inside the study area; 3) the clocks which assign the time stamp on each transponder read always need to be kept synchronized, otherwise this system will suffer from clock-biased problem; 4) this system utilizes unique tag IDs that correspond to individual drivers of probe vehicles which will elevate privacy issues; and 5) this system needs a large amount of data storage space.

2.3.3 Ground-Based Radio Navigation

This probe vehicle system works by using a receiving antenna network and probe vehicles equipped with electronic transponders. The process of ground-based radio navigation

communication is illustrated in Figure 3. This system is commonly used by transit agencies and private companies to manage fleet operations.

Compared to the other probe vehicle systems, the advantages of this system are 1) the initial cost needed to run this system is low; 2) the more people using this system, the more widespread the use of the technology; and 3) if a ground-based location provider service is available, the data collection is relatively simple.



Source : [1]

Figure 3. Ground-based radio navigation system

This system has some disadvantages, too: 1) it suffers from low accuracy since the precision is affected by the topography of the land and mounted in-vehicle equipment; 2) the technology is

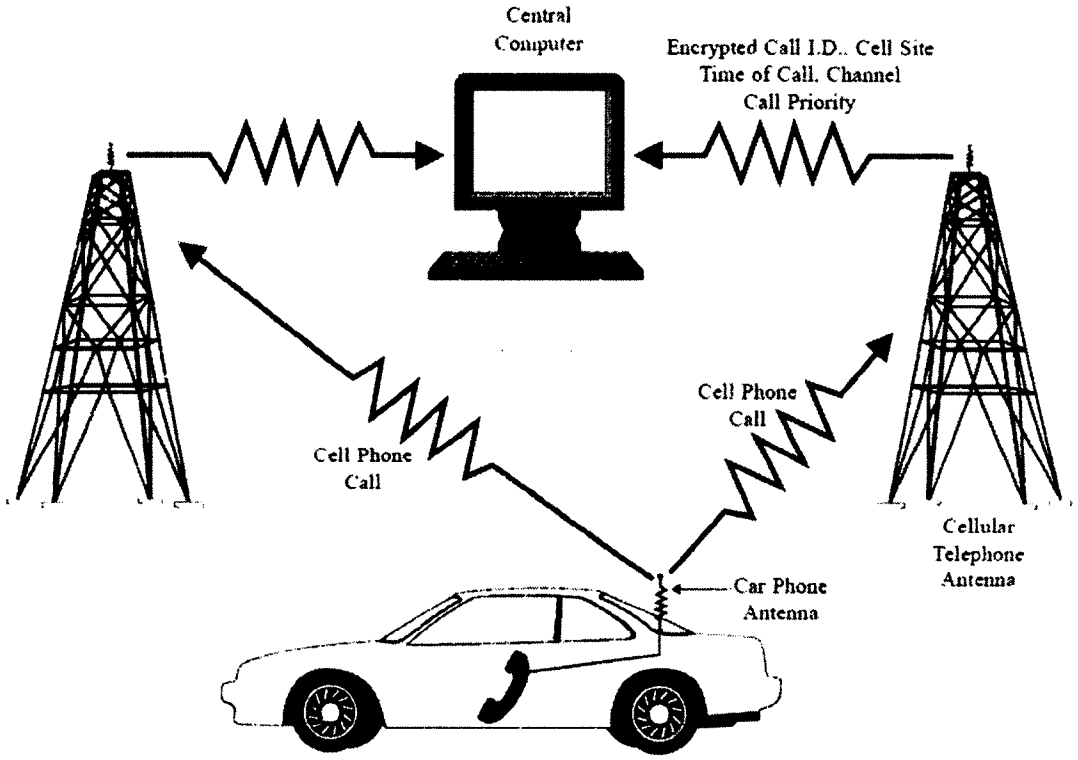
outdated and less precise compared to GPS technologies; 3) the density of urban areas will affect signal penetration; 4) since this technology is used by commercial vehicles and transit agencies, the data may not represent the overall driving population and may be biased; 5) as this system is mainly operated by private companies, it will require their approval to get the data; 6) this system is currently available in certain cities particularly in dense urban areas with many transit agencies; and 7) this system will need driver recruitment which depends on the size and scope of the study.

2.3.4 Cellular Phone Tracking

There are two techniques to collect data for this probe vehicle system: cellular telephone reporting and cellular geolocating. In the cellular telephone reporting, the probe vehicle driver needs to call an operator of the central facility to inform that he or she has passed a certain checkpoint. The operator will then record the driver's identification, location, and time of the call. The data between reporting locations can be determined by evaluating the time between successive telephone calls. This method is very useful in evaluating traffic conditions particularly when the traffic experiences delay after an unexpected event (e.g. incidents). The other technique of this system, cellular geolocating, was conducted only once in an operational test in the Washington, D.C. area, sponsored by private and public organizations under the project name Cellular Applied to ITS Tracking And Location (CAPITAL).

As this technique utilizes an existing cellular telephone network, vehicle locating devices and a central control facility to collect data, all vehicles equipped with cellular telephones are potential probe vehicles. Every activity of cellular phones from vehicles is monitored by the system. The illustration of this system is shown in Figure 4. The advantages of this system compared to other probe vehicles systems are 1) no driver recruitment is necessary since all the vehicles with

cellular phones can be potential probe vehicles; 2) this system does not need in-vehicle equipment to run and 3) as all vehicles with cellular phones can be probe vehicles, this system has a large potential sample size.



Source :[1]

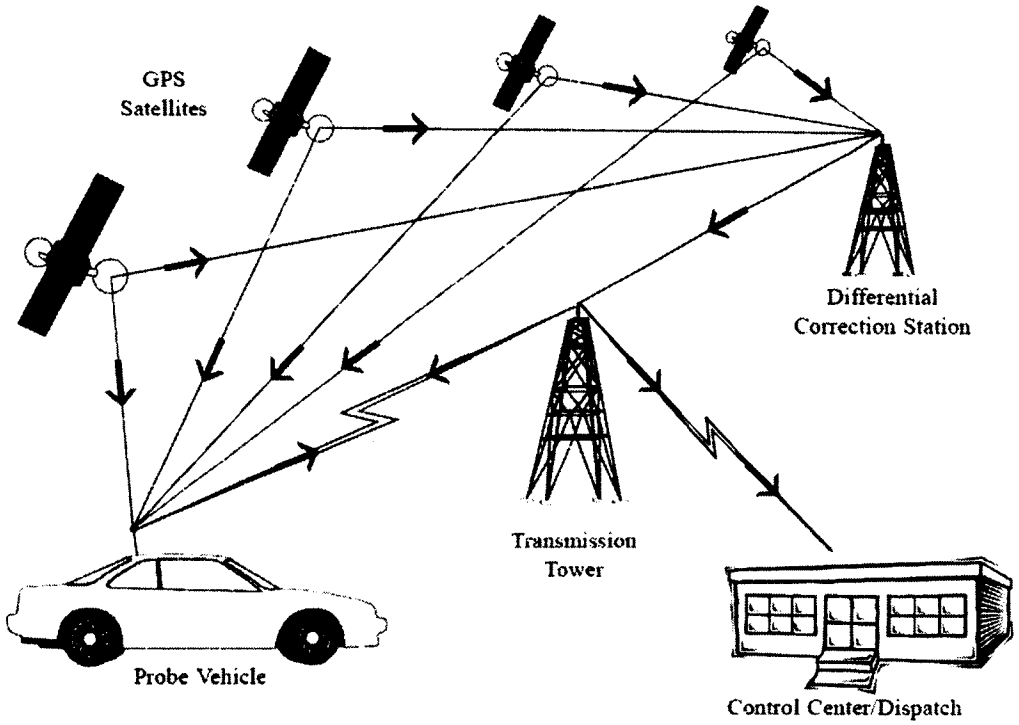
Figure 4. Cellular Geolocation System for probe vehicles

The disadvantages of this system are, 1) until now, the cellular geolocating has been conducted only once through the CAPITAL Operational Test in the Washington, D.C. area; 2) this system has a high potential for privacy issues; 3) this system depends very much on infrastructure (in this case cellular infrastructure) and thus the data collection is limited to roads within the coverage area of the cellular network; 4) proper use of this system really depends on cellular

phone use, and consequently data collection can break down during low cellular telephone use; 5) this system suffers from inaccurate data collection due to topography and line of sight barriers on the geolocating process; and 6) this system only represents the driving behavior of motorists that use cellular telephones while driving, and this sample does not necessarily represent the whole population of motorists in the study area, thus this system suffers from potentially biased samples.

2.3.5 Global Positioning System (GPS)

The Global Positioning System (GPS) is a space-based navigation system using satellite that gives location and time information anywhere on or near the earth in all weather conditions.



Source : [1]

Figure 5. Global positioning system for probe vehicles

A GPS works properly only when there is unobstructed line of sight to four or more GPS satellites. The system gives significant assistance to military, civil, and commercial users around the world. The system is operated and maintained by the United States government and can be accessed freely by anyone around the world with a GPS receiver.

The system was originally built by the U.S. Department of Defense for tracking military ships, aircraft and ground vehicles. The signals are sent from 24 satellites orbiting the earth at 20,120 km (12,500 mi). These signals are capable of providing location, direction, and speed anywhere in the world, which has become the main reason for utilizing this system for probe vehicles. Figure 5 illustrates the communication process of the GPS probe vehicle system.

Compared to other probe vehicle systems, the GPS probe vehicle technique benefits from these advantages: 1) relatively low operating cost after initial installation; 2) capability of providing data continuously for all time and space; 3) availability is improving continually; and 4) using the GPS receiver, the data are collected automatically. However, this system also suffers from these disadvantages: 1) privacy issues arise as this system provides the location and time of the driver anywhere in the world; 2) large buildings, trees, tunnels, or parking garages can impair the signal, resulting in data loss; 3) the difference in driving behavior makes it hard to achieve consistency between drivers, 4) two-way communication systems are needed to make this system work; and 5) this system has a relatively high installation cost, mainly for hardware investment.

2.4 Probe Vehicle Sample Size Issue

An essential issue in using the probe vehicle system is sample size. The question is, out of the entire vehicle population, what is the probe vehicle percentage needed for obtaining reliable data? According to [21] in a project called ADVANCE (Advanced Driver and Vehicle Advisory

Navigation ConcEpt), for a simulation model of the city of Chicago that has a total of 3,946 links over 465 square miles, 4,000 vehicles would ideally be needed as probe vehicles for a 200 square mile test area.

A research study [22] conducted on this topic for travel time data collection found that the number of probe vehicles increases non-linearly as the reliability criterion is made more stringent. Another study [23] also tried to explore this topic for arterial speed estimation and found that their network needs to have 4% to 5% of active probe vehicles, or 10 vehicles minimum, passing through a link in the sampling period.

2.5 Previous Study in Probe Vehicle System Usage

The topic of probe vehicle systems has been explored extensively. In this section, several research papers that have been conducted on this topic will be discussed to investigate their relevance to this study.

2.5.1 Probe Vehicle System to estimate travel time

There has been a lot of research conducted to evaluate the application of probe vehicle systems in estimating travel time at all levels (microscopic, mesoscopic, and macroscopic) of the transportation network. For example, by setting up experiments using a micro simulation model, [24] probe vehicle data from a signalized arterial were used to calibrate a delay (travel time) distribution for an isolated intersection that was derived from an analytical model under different circumstances. The authors concluded that the delay distribution, estimated based on both least-squares and maximum likelihood method under sparse traffic conditions, can well represent the simulation data. However, when the degree of saturation increases, the maximum likelihood method performs better than least-squares method. In another study, [11] event-based traffic data

were used to estimate time-dependent travel time by tracing a virtual probe vehicle and determining its maneuver (acceleration, deceleration and no speed change) based on estimated traffic states. The authors claimed that the proposed model could generate accurate time-dependent travel times under various traffic conditions. A study for estimating travel time using a probe vehicle system at the macroscopic level was conducted by [25]. They utilized a statistical model (maximum likelihood estimation) to estimate urban road network travel time for any route between two points under specified trip conditions by using vehicle trajectories obtained from low frequency GPS probes as observations. They showed that sparse probe vehicle data can potentially monitor the performance of the urban transport system. Another way for evaluating the travel time with probe vehicle system is by using Adaptive Kalman filter as described by [4] who performed travel time estimation for a freeway. Using a microscopic simulation model, the authors improved travel time estimates by incorporating data from a small sample of probe vehicles and proposed an Adaptive Kalman Filter-based method that can dynamically estimate noise statistics of a system model by adapting to real-time data. They discovered that the proposed algorithm significantly enhanced section travel time estimates compared to the cases when a single data source was used.

2.5.2 Probe Vehicle System to estimate traffic states

Travel time information is crucial to determine traffic states, thus by determining the travel time information one also can identify the traffic condition as shown by [26]. However, some research about traffic states determination using probe vehicles has been conducted by observing other traffic parameters and evaluating the travel time indirectly. For example, [27] conducted a study to estimate freeway density with both loop detector data and IntelliDrive-based probe vehicle data. This study used probe vehicle data to help determine density distribution between two loop

stations. The authors found that the proposed method is valid and the algorithm can be used offline and in real time. Based on probe and loop detector data [28], others proposed an empirical model of the effect of sample size and detector spacing on the accuracy of freeway congestion monitoring in estimating delay, average duration of congestion and average spatial extent of congestion. The model facilitates comparison of the effect of sample size and detector spacing in regard to the cost for achieving the same level of accuracy. The authors showed that the result could be used as a guide to determine the sample size or detector spacing in planning new congestion monitoring. At a regional level, [29] proposed a probabilistic modeling framework for estimating and predicting arterial travel time distributions using sparsely observed probe vehicles. Their model was based on hydrodynamic traffic theory to learn the density of vehicles on arterial road segments illustrating the distribution of delay within a road segment. They utilized 500 taxis as probes to test the method to estimate real-time traffic for a subnetwork of San Francisco.

2.5.2.1 Probe Vehicle System for automatic incident detection and weather condition detection

Another application of probe vehicle system is to detect incidents automatically as demonstrated by [6], where the researchers developed a neural network model for automatic incident detection using simulated data derived from inductive loop detectors and probe vehicles. They showed that it is feasible to develop advance data fusion neural network architectures or detection of incidents on urban arterials using data from existing loop detector configurations and probe vehicles. By the use of microsimulation for generating probe data, [30] investigated the use of vehicle-infrastructure integration, that is, vehicles equipped to collect traffic probe data that can be used to assist transportation planning or operations. Considering different data requirements, in their work the authors considered application to weather condition detection, incident

detection, and real-time adaptive signal control where in each application they evaluated the ability of the probe sampling system to support the application. They found that the least-demanding application (such as weather condition detection), could be served well at low probe vehicle market penetration rates, but the most demanding applications (real-time adaptive signal control) would need the majority of the vehicles to be equipped.

2.5.3 Probe Vehicle System to estimate queue length

The first attempt to estimate queue length using vehicle probe data was done by [5] where they used a statistical method to build an analytical formulation based on conditional probability distribution to estimate the real time queue lengths and their variance from probe vehicle information in a queue at an isolated and undersaturated intersection. They found that in the case of steady-state conditions and known arrival rates, the queue length can be estimated using only the location of the last probe vehicle in the queue. Since vehicle sample size is the main problem of probe vehicle system usage, their study also considered this issue as one of its discussion topics. From results obtained for undersaturated conditions, they came to the conclusion that the queue lengths can be estimated more effectively at relatively high volumes.

Later, in a study about estimating queue dynamics at signalized intersections, [15] introduced a method to estimate the LWR shockwave profile by determining the critical point using probe vehicle data. The critical points were determined from the time and space coordinates (vehicle trajectories) at the events when probe vehicles join the back of the queue. By knowing this critical point, the queue dynamics at traffic signals can be estimated. This method was built from, and thus applied to, the case of oversaturated intersections where queue overflow occurred at most signal cycles. Another study about estimating queues in an urban network was conducted by [17]. This study proposed a method using probe GPS data to identify the queue profile in the

time-space plane that forms the shockwave profile from which the queue dynamics can then be estimated. Since the shockwave profile was formed by classifying and clustering the probe data, the proposed method does not require any explicit information about signal setting and arrival distribution, and thus they claimed this method can be applied both in undersaturated and oversaturated intersections.

2.5.4 Probe Vehicle System to estimate emissions

At the microscopic level, vehicle fuel consumption and emissions were estimated based on instantaneous speed and acceleration level as shown by [31]. The instantaneous speed and acceleration can be obtained from vehicles trajectories. An attempt to estimate vehicle emissions at the microscopic level using vehicle trajectories from probe vehicle data has been demonstrated by [8], where they estimated the total fuel consumption and CO_2 emissions at a signalized intersection from the probe vehicle data by employing several methods, including simple extrapolation and using trajectories of two consecutive probe vehicles for each signal cycle. In this latter method, the average of two fuel consumption values (one for each trajectory) is taken to estimate the fuel consumption for the non-probe vehicles. In this study, fuel consumption and emissions were estimated using Virginia Tech Comprehensive Power-Based Fuel Consumption Model (VT-CPFM). An interesting conclusion from this study is the information about vehicle type and brand will not enhance the estimation accuracy significantly when a proper vehicle representation is introduced. This study was done by comparing every scenario to the "ground truth" of probe vehicle data attained from a microsimulation model. This study was one of a few studies for evaluating probe vehicle penetration that conducted using a microsimulation model. The reason for using the simulation model is because it is very expensive to obtain a ground truth for probe vehicle data in real life. However, since the vehicle trajectories were generated by

assigning a mean value, is it reliable to estimate fuel consumption from vehicle trajectories generated by a microsimulation model? This question was the main study topic used by [32] to examine the relevance of using vehicle trajectories simulated by dynamic traffic models for estimating fuel consumption. Their study focused mainly on the feasibility of a microscopic traffic model for fuel consumption estimation. Fuel consumption, which is estimated for real and simplified trajectories, and also connections between kinematic and the fuel consumption error, were investigated. The results showed that simplifying trajectories causes fuel consumption underestimation. A contribution of their study was a method developed to quantify and reduce the errors occurring at each kinematic phase when acceleration distributions are approximated by their mean values.

2.6 Shockwave Profile

This current study employed a shockwave profile to predict the trajectories for all non-probe vehicles. The shockwave profile was predicted using vehicle trajectories based on the principal of Kinematic Wave Theory (also known as LWR) macroscopic model. This section will describe the shockwave profile and traffic conditions in the signalized intersection.

2.6.1 The shockwave profile at signalized intersection

The first dynamic traffic model that used the fundamental diagram in the traffic conservation law was proposed by Lighthill and Whitham [33] and Richards [34], and is known as the LWR model. This traffic flow model was built based on the fluid dynamics continuity equation, known as the first-order LWR traffic flow model. Using a traffic flow fundamental diagram, the general LWR shockwave model can be built.

Figure 6(a) shows the fundamental diagram and the shockwave speed due to the interrupted traffic flow during red light at signalized intersection. As the light turns red (point *A*), the capacity will drop to point *B* (capacity = 0). Point *C* represents the capacity of the link.

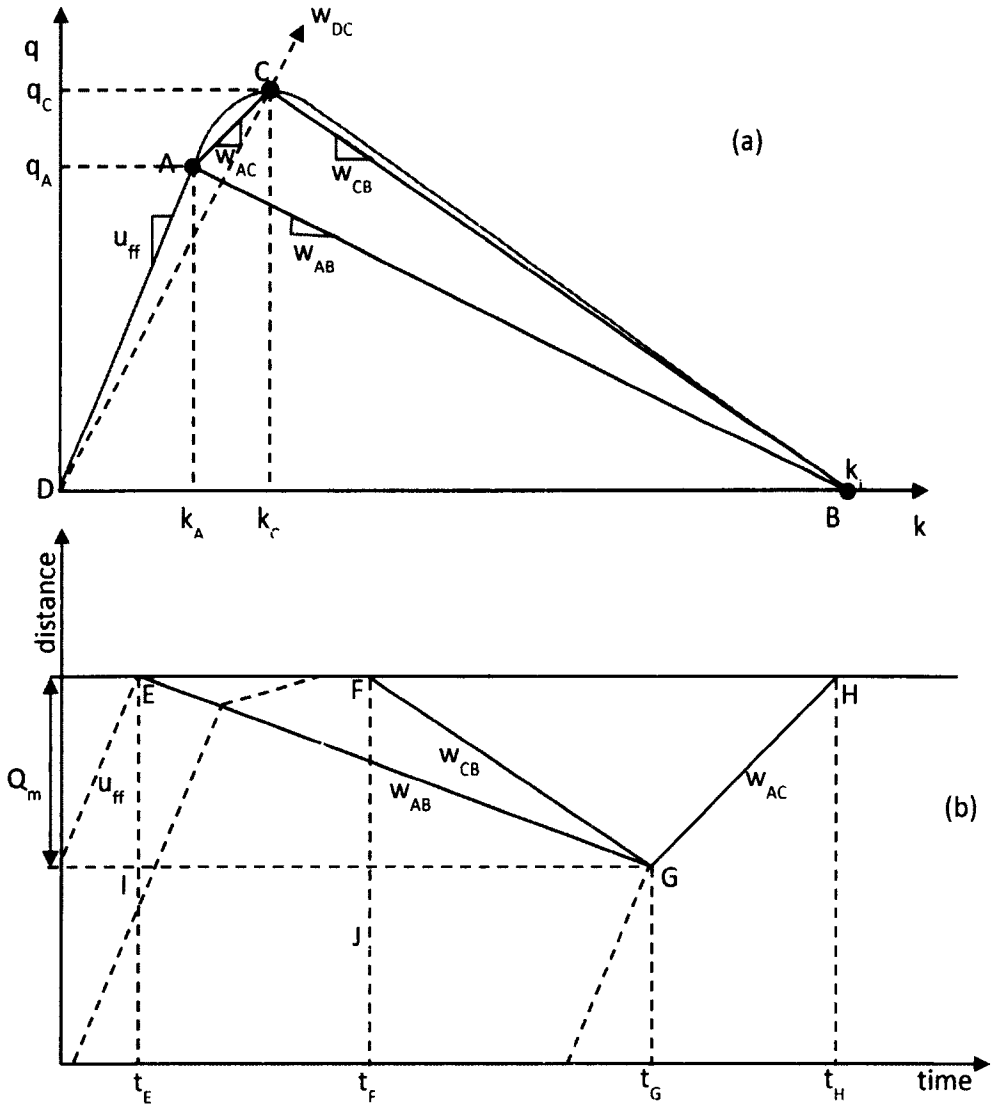


Figure 6. The fundamental diagram and incidents shockwave profile

Figure 6(b) shows the shockwave profile of the link due to an incident. Let us assume r denotes the duration of the red time. In Figure 6, $r = t_F$-t_E, while t_G-t_F is defined as the total time from the beginning of green to the time of last vehicle joining the queue, t_H-t_F is the total time from the beginning of green to the condition where the traffic reaches the free flow speed and Q_m is the queue length. The total delay (TD) is the sum of the areas of triangles *EFG* and *FHG* multiplied by the density of each associate traffic state.

Using the fundamental diagram, the shockwave formation in time and space is drawn (see Figure 6). The shockwave speed w_{AB} is the queue formation speed as the queue propagates upstream when the traffic is stopped for the red light, thus w_{AB} is defined as queuing shockwave speed, calculated as equation (1).

$$w_{AB} = \frac{q_B - q_A}{k_B - k_A} \tag{1}$$

As the lights turn green, vehicles begin to discharge at saturation flow rate, forming another shockwave with shockwave speed w_{CB} moving upstream, this shockwave is defined as discharge shockwave speed, formulated as equation (2)

$$w_{CB} = \frac{q_B - q_C}{k_B - k_C} \tag{2}$$

The position of the last vehicle in queue defines the maximum queue length. At this point the queuing and discharge shockwave meet and a third shockwave with shock wave speed w_{AC} is generated propagating downstream. This shockwave is formed when the vehicle from upstream reaches the vehicles who are just leaving the queue with speed at capacity. This shockwave speed is defined as departure shockwave, shown by equation (3)

$$w_{AC} = \frac{q_C - q_A}{k_C - k_A} \tag{3}$$

With the known shockwave speed w_{AB} and w_{CB} and the length of lane closure duration (r), by utilizing ΔEGI and ΔFGJ in Figure 6, yields

$$w_{CB} = \frac{Q_m}{t_G - t_F} \tag{4}$$

$$w_{AB} = \frac{Q_m}{r + (t_G - t_F)} \tag{5}$$

By equating the Q_m from equations (4) and (5), the total time from the beginning of green light to the time of the last vehicle dissipating from the queue, t_G-t_F (minutes) can be determined as shown by equation (6),

$$t_G - t_F = \frac{w_{AB} \cdot r}{w_{CB} - w_{AB}} \tag{6}$$

The queue length can be determined by the following procedure:

$$Q_m = w_{CB} \cdot t_G - t_F \tag{7}$$

Substituting $t_G - t_F$ from the equation (6) into equation (7), the queue length (kilometers) can be determined as in equation (8),

$$Q_m = \frac{r}{60} \frac{|w_{CB}| \cdot |w_{AB}|}{|w_{CB}| - |w_{AB}|} \tag{8}$$

Using t_G-t_F and Q_m the total time from lane opening to normal conditions is formulated as follow, as in equation (9)

$$t_H - t_G = \frac{Q_m}{w_{AC}} \tag{9}$$

and in equation (10)

$$t_H - t_F = (t_H - t_G) + (t_{G-}t_F) \tag{10}$$

Thus, the time needed from the beginning of green back to free-flow traffic condition can be obtained as in equation (11),

$$t_H - t_F = \frac{Q_m}{w_{AC}} + (t_{G-}t_F) \tag{11}$$

As stated earlier, total delay due to the red light interruption is the area of ΔEFG plus the area of ΔFHG multiplied by each traffic state density, as shown by equation (12).

$$TD = \frac{r \,.\, Q_M}{2}(k_B - k_A) + \frac{(t_H - t_F) \,.\, Q_M}{2} + (k_C - k_A) \tag{12}$$

2.6.2 Traffic flow conditions in the signalized intersection

There are two kinds of conditions occurring in signalized intersections: undersaturated and oversaturated. The undersaturated condition is the situation where the green time can accommodate all the queue and incoming traffic and thus clear all the queued vehicles in one cycle. On the other hand, the oversaturated condition is the situation where there is overflow queue (also known as spillover) where the queued vehicles will not be able to pass the intersection in one cycle. This study will focus on the undersaturated traffic flow condition and leave the oversaturated one for future work.

2.7 Emissions Model

Similar to transportation models, emissions models have been built at the macroscopic, microscopic, and (recently) mesoscopic level. The macroscopic model uses average aggregate parameters such as average speed or mean speed distribution to estimate network-wide fuel energy and emissions rates. In average speed models, fuel consumption rates are a function of trip time, trip distance, and average speed. According to [35], as cited in [36], since macroscopic models did not take into account aerodynamic drag resistance at high speed, they could only be used appropriately for average speeds of less than 50 km/h. At average speed over 55 km/h, as shown by [37], the aerodynamic effects on fuel consumption become significant.

Microscopic models, on the other hand, estimate second-by-second vehicle fuel consumption and emissions rate which are aggregated to estimate network-wide emissions rates [38]. The microscopic emissions model can be implemented to evaluate emissions from specific driving cycles or integrated directly with microscale traffic simulation (e.g. Vissim, Transmodeller, Paramics, etc). However, the problem with this approach as discussed in [39] is that it is more complicated to use this model for estimating larger, regional emissions because microscale models typically require extensive data on the system of the study location and are restricted in size due to the non-linear complexity that occurs within larger networks.

The mesoscopic emissions model was created to overcome this problem. This model, as demonstrated by [36], employed link-by-link parameters such as average travel speed, number of vehicle stops per unit distance and average stop duration to construct synthetic drive cycles for the roadway segments. Those drive cycles were treated as inputs in the microscopic model to calculate emissions.

There are several known emissions models; at the macroscopic level there are the Elemental model and Watson model which were developed using average speed. At the microscopic level, there are CMEM and VT-MICRO models which utilize instantaneous speed and acceleration to predict the pollutant. Several efforts have been done to model the fuel consumption and emissions modeling at themesoscopic level, namely the Akcelik Model, MEASURE model and VT-Meso model. In the U.S., previously state and local agencies for transportation planning used a model developed by EPA called MOBILE. The latest version of this is MOBILE 6.2. This is a macroscopic emissions model; recently, the EPA released a new model called MOVES to replace the MOBILE model that can estimate emissions in macroscopic, microscopic, and even mesoscopic level.

2.7.1 Elemental Model

This model, proposed by Herman and colleagues [37, 40], used average speed for its macroscopic emissions model. This model stated fuel consumption in urban areas as a linear function of the average trip time per unit distance, formulated as shown by equation (13)

$$\Phi = K_1 + K_2 T, V < 55\ km/hr \qquad (13)$$

where Φ = fuel consumption per unit distance

T = average travel time per unit distance, and

$V(=1/T)$ = average speed

K_1 (mL/km) represents the vehicle mass, while K_2 (mL/sec) is a function of vehicle average speed.

2.7.2 Watson Model

The Watson model [41], as in [36], used average speed to develop the fuel consumption model. It integrated the changes in the positive kinetic energy during acceleration as a predictor variable, formulated as shown by equation (14)

$$F = K_1 + \frac{K_2}{V_s} + K_3 V_s + K_4 PKE \tag{14}$$

where F = fuel consumed (L/km)

V_s = space mean speed (km/hr)

PKE represents the sum of the positive kinetic energy changes in acceleration process (m/s^2) and is expressed in equation (15),

$$PKE = \sum \frac{V_f^2 - V_i^2}{12.960 X_s} \tag{15}$$

where V_f = final speed (km/hr) f

V_i = initial speed (km/hr) i

X_s = total section length (km)

2.7.3 CMEM

CMEM stands for Comprehensive Modal Emissions Model [39], and is a fuel consumption and emissions model built by the College of Engineering-Center for Environmental Research and Technology (CE-CERT) at the University of California-Riverside with researchers from the University of Michigan and the Lawrence Berkeley National Laboratory in a four-year research project sponsored by the National Cooperative Highway Research Program (NCHRP, Project 25-

11). CMEM uses a physical power-demand modal modeling approach based on a parameterized analytical representation of emissions production. The instantaneous emissions were modeled as a product of fuel rate (FR), engine-out emission indexes ($g_{emission}/g_{fuel}$), and catalyst pass fraction (CPF) as shown by equation (16),

$$tailpipe\ emissions = FR * \left(\frac{g_{emission}}{g_{fuel}}\right) * CPF \qquad (16)$$

where : FR = fuel-use rate in grams/s,

$g_{emission}/g_{fuel}$ = grams of engine-out emissions per grams of fuel consumed,

CPF = the catalyst pass fraction, defined as the ratio of tailpipe to engine-out emissions

The modal emissions model consists of six modules: engine power demand, engine speed, air-fuel ratio, fuel rate, engine-out emissions and catalyst pass fraction.

2.7.3.1 Engine Power Demand Module

The engine power demand module was formulated as in equation (17)

$$P_{tract} = A.v + B.v^2 + C.v^3 + M.a + M.g.v.\sin\theta$$

$$P = \frac{P_{tract}}{\eta_{tf}} + P_{acc} \qquad (17)$$

where, P_{tract} = total tractive power (kw)

A = coefficient of rolling resistance

B = coefficient of speed-correction to rolling resistance

C = coefficient of air-drag factor

v = speed (m/sec)

a = acceleration (m/s^2)

g = the gravitational constant (9.81 m/s^2)

θ = the road grade angle

P = the engine power output

η_{tf} = the combined efficiency of the transmission and final drive

P_{acc} = the engine power demand associated with the operation of vehicle accessories such as air conditioning, power steering and brakes, and electrical loads.

2.7.3.2 Engine Speed Module

In this module, engine speed is simply expressed in vehicle speed, using gear ratios and a shift schedule to determine upshift or downshift. In the air/fuel ratio module, the air/fuel ratio is from three regions: lean, stoichiometric and rich.

2.7.3.3 Fuel Rate Module

The Fuel Rate Module was formulated as in equation (18)

$$FR \approx \emptyset\left(kNV + \frac{P}{\eta}\right)\frac{1}{44} \tag{18}$$

where, k = the engine friction factor,

N = engine speed (revolutions per second)

V = engine displacement (liter),

$\eta \approx 0.4$ = a measure of indicated efficiency

2.7.3.4 Engine-out emissions Module

The engine-out emission module was formulated as in equations (19) to (22)

$$ECO \approx [C_0(1-\emptyset^{-1}) + a_{CO}]FR \quad (19)$$

$$EHC \approx a_{HC}FR + \gamma_{HC} \quad (20)$$

$$ENO_x = a_{1NOx}(FR - FR_{NOx})\emptyset < 1.05 \quad (21)$$

$$ENO_x = a_{2NOx}(FR - FR_{NOx})\emptyset \geq 1.05 \quad (22)$$

2.7.3.5 Catalyst Pass Fraction Module

The catalyst pass fraction module was formulated as in equation (23)

$$CPF(ei) = 1 - \varepsilon_{ei}.\exp\{[-b_{ei} - c_{ei} * (1-\emptyset^{-1})] * FR\} \quad (23)$$

where, ei = either CO or HC emissions

ε_{ei} = the maximum catalyst CO or HC efficiency

FR = the fuel rate (grams/second)

B_{ei} = the stoichiometric CPF coefficients

c_{ei} = the enrichment CPF coefficient

2.7.4 VT-MICRO

The Virginia Tech Microscopic Energy and Emission Model (VT-Micro Model) [31, 42, 43] was developed for estimating vehicle fuel consumption and emissions rate which included CO, HC and NO_x for five light-duty vehicles and three light-duty trucks as a function of vehicle

instantaneous speed and acceleration level. The data were collected at the Oak Ridge National Laboratory (ORNL). The VT-Micro models estimate the instantaneous fuel consumption and emissions rates of individual vehicles have a general formula as shown in equation (24),

$$MOE_e = \begin{cases} \sum_{i=0}^{3}\sum_{j=0}^{3} \exp\left(k_{i,j}^{e}.v^{i}.a^{j}\right) \text{ for a } \geq 0 \\ \sum_{i=0}^{3}\sum_{j=0}^{3} \exp\left(l_{i,j}^{e}.v^{i}.a^{j}\right) \text{ for a } < 0 \end{cases} \quad (24)$$

where : MOE_e = Instantaneous fuel consumption or emissions rate (L/s or mg/s),

a = Instantaneous acceleration of vehicle (km/h/s)

v = Instantaneous speed of vehicle (km/h),

$k_{i,j}^{e}$ = Vehicle-specific acceleration regression coefficients for MOE_e

$l_{i,j}^{e}$ = Vehicle-specific deceleration regression coefficients for MOE_e

2.7.5 MOBILE and MOVES

MOVES is the latest regulatory mobile emissions model released by the U.S. Environmental Protection Agency to replace the former emissions model called MOBILE. MOVES was designed to estimate the vehicle emissions accurately under a wide range of user-defined conditions [44]. Table 3 exhibits the comparison between MOBILE 6.2 and MOVES2010.

Table 3. MOBILE 6.2 and MOVES2010 comparison

Criterion	MOBILE 6.2	MOVES2010
Model methodology	Based on average speed	Based on modal activity
Software interface	Model embedded calculation	Graphical user interface
Emissions source	On-road	On-road and off-road
Spatial scale	Single large regional scale	Three scales of analysis: macroscopic, mesoscopic, and microscopic
Pollutants	Criteria pollutants, hydrocarbons, particulate matter, air toxics, GHGs-CO2, methane	All pollutants plus new ones: Sulfur dioxide (SO_2), ammonia (NH_3), nitrogen oxides (NO_2, NO), energy consumption
Emissions process	Running exhaust Start exhaust Hot Soak Diurnal Resting loss Running loss Crankcase Refueling Brake wear Tire wear	Running exhaust Start exhaust Extended idling Off-gassing (well-to-pump) Evaporative fuel permeation Evaporative fuel vapor venting Evaporative fuel leaking Brake wear Tire wear
Roadway classification	Freeway Arterial and collector roads Local Freeway on- and off-ramps	Rural restricted access Rural unrestricted access Urban restricted access Urban unrestricted access
Vehicle classification	28 vehicle classes	13 vehicle classes
Fuel type	Gasoline Diesel Compressed natural gas	Gasoline Diesel fuel Compressed natural gas (CNG) Liquid propane gas (LPG) Ethanol (E85) Methanol (M85) Gaseous hydrogen Liquid hydrogen Electricity

Table 3. (cont.)

Temporal scale	Analysis years: 1952 - 2050	Analysis years: 1999 – 2050
Speed	Single speed for ramps and local roads	Speed distribution for all roadway types by area type (urban or rural)
Emissions estimation	Trip-based vehicle average speed	Distributes total activity into source and operating mode bins
Meteorology data	User supplied	Default county-specific temperature and humidity values; users can overwrite the default data with local specific data
Fuel supply	User supplied	Default county-specific fuel supply values; users can overwrite the default data with local specific data
Inspection and maintenance program	User supplied	Default county-specific inspection and maintenance program values; users can overwrite the default data with local specific data
Age distribution	User supplied – registration distribution	Default national age distribution for years 1999-2050
Output	Emissions factors	Emissions inventories or emissions factors, total energy consumption
Other significant features	None	Ability to analyze advanced technology vehicles (e.g. hybrid vehicles)
		Modal-based
		Converters to translate MOBILE6 inputs to MOVES

Source : modified from [45]

2.7.6 Akcelik Model

In the fuel consumption model proposed by Akcelik [46], three portions of the urban driving cycle which are cruising, idling and the deceleration-acceleration cycle were separated and then the fuel consumptions of each portion were estimated separately. The Akcelik model was formulated as in equation (25),

$$F = f_1 X_s + f_2 d_s + f_3 h \tag{25}$$

where F = average fuel consumption per roadway section (mL)

X_s := total section distance (km)

d_s = average stopped delay per vehicle (secs)

h = average number of stops per vehicle

f_1 = fuel consumption rate while cruising (mL/km)

f_2 = fuel consumption rate while idling (mL/sec)

f_3 = excess fuel consumption per vehicle stop (mL).

2.7.7 MEASURE Model

MEASURE (Mobile Emission Assessment System for Urban and Regional Evaluation) was a modal emissions model developed by researchers at the Georgia Institute of Technology. This model is a GIS-based emissions model that predicts different vehicle modes and produces mesoscopic estimates of HC, CO and NOx [47, 48]. MEASURE was designed to suit the traditional four-step travel demand modeling. It consists of start emissions module and on-road emission module.

2.7.8 VT-Meso Model

Similar to the VT-Micro Model, the Virginia Tech Mesoscopic Energy and Emission Model (VT-Meso Model) was developed at Virginia Tech. This model was proposed by Yue and Rakha and documented in Yue's dissertation [36]. It was built using the microscopic vehicle fuel consumption and emissions model that was developed earlier at Virginia Tech (VT-Micro). In this mesoscopic model, link-by-link input parameters which are average travel speed, average number of stops per unit distance and average stop duration were employed to construct a synthetic drive cycle and compute average link fuel consumption and emissions rates. The model then estimates the proportion of time that a vehicle typically spends for cruising, decelerating, idling and accelerating when it traversed a link. Using the synthetic drive cycle and the time proportion, fuel consumption and emissions were estimated by utilizing VT-Micro model where several models were employed for each mode of operation. Afterwards, the total estimation along a segment is obtained by summing across the different modes and dividing by the distance traveled. The VT-Meso model was built for normal and high emitting vehicles.

CHAPTER 3

LANE IDENTIFICATION

3.1 Introduction

Instrumented vehicles, known as probe vehicles, have become increasingly popular since the current Intelligent Transportation Systems (ITS) strategies necessitate real-time traffic information. Among several probe vehicle systems, the Global Positioning System (GPS) probe vehicle systems has been recognized as a very efficient way to collect online data, especially at the network level [1]. This system has become more feasible since most portable devices such as smartphones and GPS navigation systems can be used as a platform to collect vehicle trajectory data. Processing large amounts of real-time data from these GPS systems reliably is a significant challenge and critical for optimizing transportation systems.

Several issues have been addressed with regard to the utilization of GPS probe vehicle data for improving the transportation systems. One of the most essential ones is the issue of GPS positioning error. It is well known that the GPS instruments are expected to produce location errors of 3 to 15 meters [1, 49, 50]. The level of error does not allow determining the probe vehicles' lane, which is crucial especially in the case of unequal queues. The current study is an attempt to predict an instrumented vehicles' lane at a signalized intersection particularly when there is a significant difference between the queue lengths of a lane as compared to the queue on another lane.

Researchers have conducted various studies to obtain travel information data using probe vehicles to estimate queue lengths particularly at signalized intersections. Several approaches have been exercised to carry out this kind of study e.g., utilizing shockwaves theory as reported

in [15, 17, 51, 52] which is the most common approach for estimating queue lengths, making use of probe vehicles' travel time as described in [53] and employing probabilistic probe vehicle based positioning method [5, 54]. Beside queue length estimation, another application of these studies was estimating vehicle's fuel consumption and emissions [8].

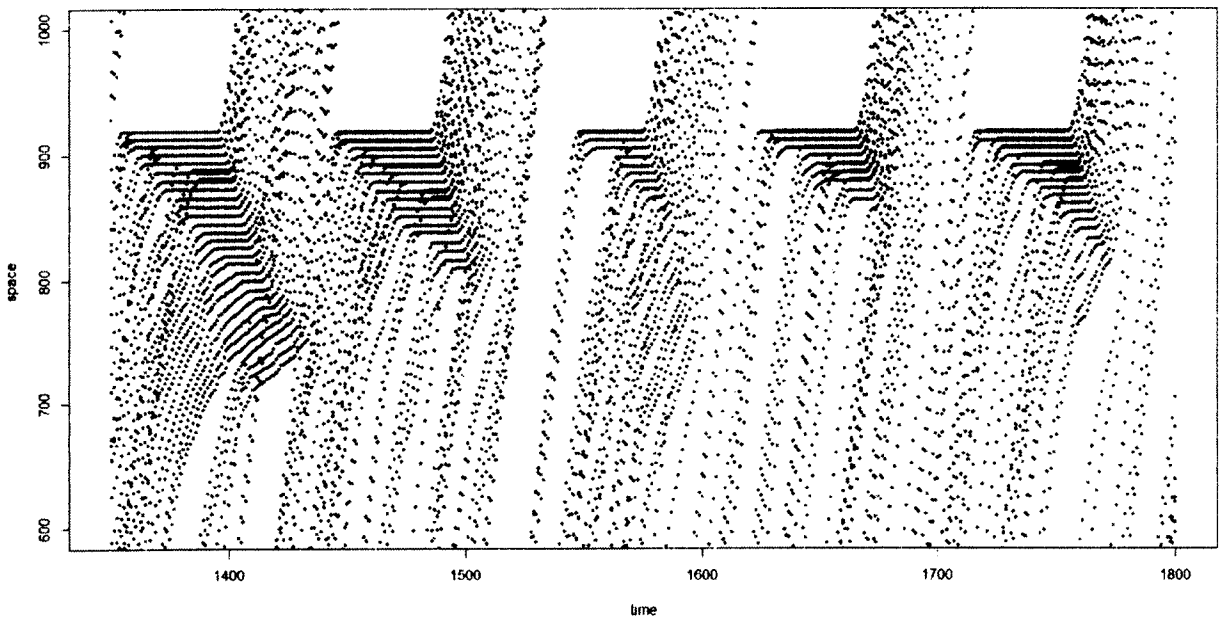


Figure 7. Vehicle trajectories show the different SW speeds of different lane

Existing probe vehicle studies, to the best of the author's knowledge, do not address prediction of the probe vehicles' lane. In the current literature, researchers either assume queuing on a single lane or equal parameters (queue length or emissions) for all lanes which may not be ideal for a real life situation. Unequal queuing, including cases where there is no vehicle at all in a lane, occur commonly at many signalized intersections. In particular, unequal queuing is normal when one of the lanes is shared (e.g., shared right-turn and through movement) or when the lanes serve traffic destined for different downstream points and demand for the lanes is unbalanced.

This dissertation is an attempt to bridge the gap in the literature by investigating different modeling approaches to infer the lane of probe vehicles. Figure 7 is an example of vehicles' trajectories of unequal queues on a two-lane intersection approach. The data for this figure were generated using PTV's transportation simulation software *Vissim* [55].

3.2 Problem Definition

Figure 7 shows there is a significant difference between the queuing shockwave speeds formed in the long queue lane and the ones formed in the short queue lane. In this study, this information was used to identify a vehicles' lane while queuing at signalized intersections.

The main objective of this study was to predict the instrumented vehicle's lane in a signalized intersection particularly in an unequal queue condition for real-time applications. This research focuses on undersaturated traffic conditions and left oversaturated conditions for future work. Several methods based on traffic flow theory and statistical learning techniques, were utilized to meet the goal. The traffic flow theory normally idealizes the traffic situations which make it prone to bias. On the other hand, probability theory normally deals well with the problem of uncertainties. This approach was chosen to deal with uncertainties related to traffic dynamic and the GPS probe system while keeping the principle of traffic theory.

3.3 Related Studies

Queue length is a crucial parameter in evaluating and optimizing traffic controls such as traffic signals. Studies have been conducted to model a queue either in an intersection or in metered on-ramps. Several approaches have been utilized or developed to do this task; the most common one is by using Lighthill and Witham [33] and Richard's [34] (LWR) shockwave theory, for example, the study by Liu et al. [56]. In their study, the intersection queue length was measured

using the queue discharge process in the immediate previous cycle. By employing the LWR shockwave theory, they were able to separate the queue discharge flow state from the upstream arrival traffic state. Vigos et al [57] employed a Kalman-Filter to estimate the number of vehicles count in signalized links based on online measurement of flow and occupancy from loop detectors. Wu et al. [58] tested three types of methods for estimating queue length at metered on-ramps: Kalman filter, linear occupancy and Highway Capacity Manual (HCM) back of queue. They concluded that the Kalman filter and linear occupancy methods are practical for real-world operations even though they have limitations, while the HCM back-of-queue method did not reliably estimate the on-ramp queue length. Cang and Su [59] attempted to predict intersection queue using optimization method. In their study, they employed the neural network models to predict the intersection queue. They claimed that starting from 3 time-steps ahead, their model was capable of providing more than 90% accuracy. Geroliminis and Skabardonis [60] modeled traffic between successive traffic signals as a two-step Markov decision pocess, while the LWR theory was used to model the traffic dynamic. Then they used this approach to estimate queue lengths and predict travel times. They claimed their model was usable for cases where the loop detector data are unknown, inaccurate, or aggregated. Another study for modeling queues at an intersection using a Markov model was conducted by Viti and Zuylen [61]. In their study, they constructed a Markov model to calculate the dynamics of the queue. They found that the model introduced in the study was suitable for solving dynamic assignment problems.

The queue length modeling studies described above were using fixed sensor data as input. One of the first research projects for queue length estimation using probe vehicle data was conducted by Comert and Cetin [5]. To estimate the expected queue length and its variance, they developed an analytical formulation based on conditional probability distribution. They also discussed the

effects of probe market penetration rate to the accuracy of their model. Using their work, one can estimate the queue length in a signalized intersection by knowing only the location information of the last probe vehicle in the queue. In a later study [54], they developed formulations to quantify the error in queue length estimation using probe vehicle data. An attempt to characterize the shockwave profile using probe vehicle data was conducted by Ramezani and Geroliminis [17]. In their paper, they constructed the queue profile by separating the input data into two groups which are stopping and moving, and then they clustered the stopped vehicles to cycles. They claimed their method was helpful for identifying spillback, constructing vehicle trajectories, and estimating fuel consumption and emissions. Beside these methods, there are several other methods that have been carried out by researchers in estimating the queue length using probe vehicles, such as identifying the critical points in a LWR shockwave profile [15] and using travel times from mobile sensors [53].

Most of the studies listed above are not lane-based estimations, except for the studies conducted by Comert and Cetin [5, 54] which, in order to be able to properly estimate the queue length, still require knowledge of the probe vehicle's lane position. As discussed before, it is problematic to determine a probe vehicle's lane position due to the GPS error. This study basically was an effort to identify the probe vehicle's lane before conducting the queue length estimation. In other words, the prediction discussed in this study becomes an input for properly estimating the queue length, particularly in a signalized intersection.

3.4 Queuing Shockwave Speeds from Probe Vehicles

The time and space coordinates at the moment the probe vehicles join the queue form a profile known as a shockwave. This back of the queue shockwave moves backward from the stop bar at a signalized intersection with a speed that is equal to the slope of the line forming the profile.

This shockwave speed is known as a queuing shockwave speed. The coordinates could be extracted from the vehicle trajectories data produced by a GPS probe vehicle. One crucial factor in determining the queue length is the arrival rate of vehicles. According to the traffic flow theory, the shockwave speed is the function of the arrival rate and thus the queue lengths of two adjacent lanes will be significantly different when the arrival rate of a lane is significantly different from one another as indicated in Figure 7.

In an ideal situation, which is normally assumed in the traffic flow theory, the queuing shockwave speed line formed from the vehicles' coordinates when they join the queue is linear. However it is also well understood that this situation generally is not the case in the real life situation. There are several factors that may invalidate this ideal situation assumption, some of which are:

- *Randomness of arrival rates*

  Because the queuing shockwave speed is the function of arrival rate, the fluctuations in the arrival rate over short time periods generate a nonlinear curve.

- *GPS error*

  Due to the GPS errors, the back of the queue (the time and space coordinates when the probe vehicles join the queue) is not linear and many times those points are not in order.

For these reasons, in this study instead of using queuing shockwave speeds generated from a group of vehicles' queue coordinates, the ones generated from individual vehicles in a cycle were used. This information is utilized to predict the vehicle's lane in the unequal queue length case. The idea is to differentiate a vehicle's lane based on its individual queuing shockwave speed. Of course, there will be variations of individual shockwave speed generated from the

same arrival rates; however, this can be approached by knowing the individual queuing shockwave speed distribution for each lane.

The aim of this study is to provide representative online traffic information, therefore the data were analyzed independently every time each cycle data became available. A linear line was drawn using the time and space coordinates of the red signal and a probe vehicle when it was joining the queue. This line slope's value is the individual queuing shockwave speed. The same method was conducted to obtain all vehicles' individual queuing shockwave speed.

The individual queuing shockwave speed is equal to a slope's value of a line that connects the time and space coordinates of an individual vehicle and red signal, calculated as shown in equation (26)

$$w_i^c = \frac{X_i^c - X}{t_i^c - t^c} \tag{26}$$

where i = 1,2,..., n; n is the number of probe vehicles in the c-th cycle; w_i^c is the individual queuing shockwave speed at the c-th cycle; X_i^c is position of an individual vehicle when it joins a queue at the c-th cycle; X is the position of the traffic signal; t_i^c is the time of an individual vehicle when it joins the queue at the c-th cycle; and t^c is the time when the signal turns red at the c-th cycle. The illustration of this shockwave speed can be seen in Figure 9.

3.5 Traffic Simulation Scenarios and Data

In this study, vehicles' trajectory data were generated from PTV's microscopic simulation software *Vissim*. A simple network consisting of a single link with two lanes was constructed as representation of an intersection arm. A traffic signal was installed downstream the link, consisting of just red and green with duration of 45 seconds for each signal (cycle time = 90

seconds). The desired speed was set to 60 km/h and lane change was not allowed. To have representative data, the simulation was run for 9,000 seconds (100 cycles).

3.5.1 Arrival rates scenarios

There are three arrival rate scenarios for unequal lane (SQL and LQL) organized in this study, i.e. 300 vph and 900 vph (scenario 1), 450 vph and 900 vph (scenario 2), and 600 vph and 900 vph (scenario 3). All the scenarios represent the undersaturated traffic flow condition.

3.5.2 Data Extraction

In this study, as stated before, the traffic flow theory and statistical model were combined to reach the goal. The essential data needed are the shockwave speed formed from the time and space coordinates when the vehicle stops and discharge at a signalized intersection. Those coordinates were extracted from the vehicle trajectory data generated in *Vissim* through the following steps (also illustrated in Figure 8):

1. Subset the trajectories for each vehicle i

$$df_i = \{df | id = i\} \tag{27}$$

where df_i = data set for vehicle i

df = the entire data set

id = vehicle index

2. Check if the vehicle is experiencing queuing or not. The queuing definition in this study is when the vehicle is in full stop (speed = 0). If it never experiences a stop, skip to the next vehicle data set.

$$\text{if } (\{df|s = 0\} = \emptyset), \{\text{next}\} \tag{28}$$

where s = vehicle speed (km/h)

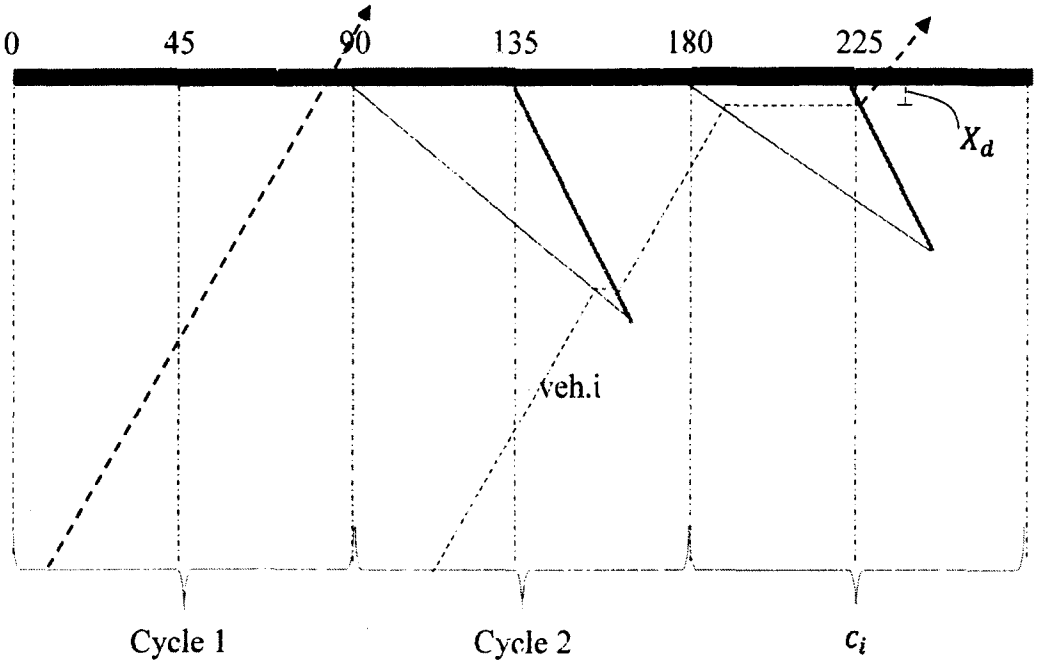


Figure 8. Trajectories of vehicle *i* in signalized intersection

3. Find the time coordinate of the first point beyond stop bar

$$t^* = \underset{t}{\text{argmin}}(X - X_0) \tag{29}$$

where t^* = time coordinate of the first point beyond the stop bar

t = time coordinate

X = distance coordinate

X_0 = traffic light position

4. Find the cycle where the first vehicle that pass the stop bar discharged

$$c_i = \frac{t^*}{CL} + 0.5 \tag{30}$$

where c_i = cycle index (rounded number)

CL = cycle length

5. Subset all reading where speed is equal to zero (vehicle was stopping)

$$S = \{df | s = 0\} \tag{31}$$

where S = data set when vehicle was stopping

6. Take the last moment of vehicle stopping (the first moment of vehicle discharge)

$$D = \{S | t = t_{max}\} \tag{32}$$

where D = data set of the moment the vehicle discharging

7. Select the vehicle stopping data only for the current cycle. This was done to avoid stopping data of vehicle i from previous cycle (see figure 9).

$$S_l = \{S \,|\, X < X_d + \alpha \,\&\, X > X_d - \alpha\} \tag{33}$$

where X_d = distance of the vehicle to the stop bar when it was discharging (taken from D dataset)

α = tolerance coefficient of vehicle movement in the queue

8. Take the arrival point data set

$$AP = \{S_l \,|\, t = t_{min}\} \tag{34}$$

where AP = arrival points (beginning of the queue) data set

t_{min} = minimum time coordinate

3.5.3 Market penetration scenarios

The probe market penetration rate was varied from 10% to 100% by increments of 10%. A vehicle was stated to be in a queue when it was completely stopped (speed = 0 km/h). The reason for this queue definition was because of the study using intersection data where the vehicle speeds are low and spacings are small.

3.6 Unsupervised Learning Methods

3.6.1 Naïve Method

The simplest way to predict the probe vehicle's lane is by clustering the individual queuing shockwave speed based on their values using a shockwave speed boundary. The next section will discuss this boundary.

3.6.1.1 The shockwave speed boundary

As explained before, individual vehicles' queuing shockwave speed can be clustered based on their lanes in the case of unequal queue where they normally will have different shockwave speed distribution. Thus, to reach this purpose, there should be a clear boundary between the shockwave speeds for each lane. However, naturally there will be shockwaves speeds that could belong to multiple lanes, especially the ones formed from the vehicles queued in the part of the road where both lanes were occupied.

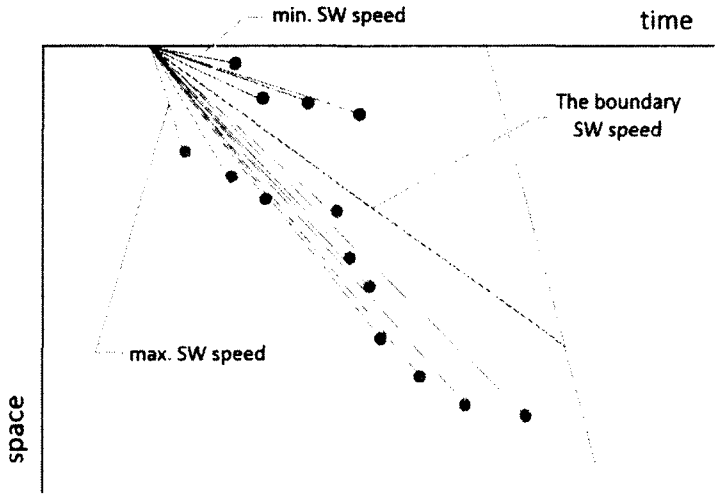


Figure 9. Illustration of individual and the boundary SW speeds

The shockwave speed boundary is a line that separates the individual queuing shockwave speeds of a lane from the other lane based on the arrival rate proportion in each cycle. In this study the boundary for two unequal lanes is computed in equation (35),

$$B_c = \min(w_c) + \left(\frac{\left(\max(w_c) - \min(w_c)\right)}{2}\right) \tag{35}$$

where B_c = shockwave speed boundary at c-th cycle;

w_c = set of individual queuing shockwave speed of c-th cycle;

Figure 9 illustrates the individual queuing shockwave speed lines and the boundary line for the two-lanes case.

3.6.1.2 Clustering

After obtaining the boundary, the probe vehicles' lanes were clustered based on their shockwave speeds relative to the boundary. The probe vehicles' lanes were identified according to equations (36) and (37).

$$if\ \min(w_c) \leq w_i \leq B_c, \quad \delta_i = 1, \quad \forall\, i \in V \qquad (36)$$

$$if\ B_c \leq w_i \leq \max(w_c), \quad \delta_i = 2, \quad \forall\, i \in V \qquad (37)$$

where δ_i = decision variable indicating the lane: $\delta_i = 1$ if vehicle on SQL, $\delta_i = 2$ if the probe vehicle on the LQL

i = index to a probe vehicle

V = the set of probe vehicles in a cycle,

3.6.1.3 Missing data cycles

Due to randomization, there could be cycles that have less than two, or even zero, probe vehicles. The number of these occurrences becomes greater as the probe vehicle market penetration rate decreases. To calculate the shockwave speed boundary, at least two probe vehicles are needed. If there are less than two vehicles, then that particular cycle will not be included in the probe vehicle lane prediction. However, it should be noted that in queue length estimation this cycle can still use the queue length estimation from the previous cycle.

3.6.2 K-Means Clustering Method

K-Means clustering is a type of hard clustering implementation where the clusters were assumed to be independent and do not overlap. In this study, the clustering was carried out for one dimension K-means clustering, i.e., taking the individual shockwave speeds and clustering them based on the lane number.

The inputs for this method are

1. Number of clusters (k)
2. Set of points, in this case the individual SW speeds (w_c) for the associated cycle
3. Partition the n observation into sets $p = \{p_1, p_2, \dots, p_k\}$. In this study, there are only two partitions ($k = 2$) which is short queue lane (p_1) and long queue lane (p_2).

The K-means objective function:

$$\arg\min_{p} \sum_{i=1}^{k} \sum_{w_i \in p_i} \|w_i - \mu_i\|^2 \tag{38}$$

where μ_i = mean of points in p_i

The K-means were conducted according to the following algorithm:

1. Place centroids $C_1 \dots C_k$
2. Repeat until convergence :
    a. For each point w_i
        i. Find nearest centroid

$$c_j = \arg\min_j D(w_i, C_j) \tag{39}$$

where $D(w_i, C_j)$ is distance between instance s_i^c and cluster center C_j.

ii. Assign the point w_i to cluster j

b. For each cluster $j = 1 \ldots k$

- A new centroid $c_j = \mu_i$ (mean of all points s_i^c assigned to cluster j in previous step)

$$C_{j_{c+1}} = \mu_i = \frac{1}{n_j}\sum_{i=1}^{n_j} w_{i_j} \tag{40}$$

3. Stop when none of the cluster assignments change (convergence).

3.7 Supervised Learning Methods

The unsupervised clustering methods discussed force at least one probe vehicle per lane. However, this may not necessarily be the case, especially for a short queue. In order to overcome this shortcoming and accommodate scenarios where probes occupy only one of the lanes, two supervised learning methods are explored in this section. Both use the optimal Bayes rule to predict the lane. The two models differ in terms of the number of variables considered. The first one is a univariate model and utilizes shockwave speeds (w_i) only and fits lognormal probability density functions to training data as explained below. The second model is bivariate and utilizes both shockwave speeds (w_i) and distance of the probe vehicle to the stop bar. The probability density functions for the second model are found by fitting a mixture of normal distributions to the training data as explained in Section 3.7.2.

3.7.1 Lognormal model

Two probability density functions fitted to the shockwave speeds (w_i) are needed for the Bayesian decision rule: one for the SQL and another for the LQL. To investigate the appropriate distributions, the shockwave speeds (w_i) from each lane were input to a probability distribution fitting software called CumFreg program [62]. The software fitted the data to the following distributions, Normal distribution. Logaritmic normal, Root normal, Square normal, Generalized normal, Logistic distribution, Logaritmic logistic, Generalized logistic, Cauchy distribution, Cauchy generalized, Exponential, Exponential mirrored, Gumbel, Gumbel generalized, Gumbel mirrored, Gumbel mirror generalized, Student's t-distribution, Weibull distribution and Weibul generalized.

The result showed that the data were distributed according to the lognormal distribution. The probability density function (pdf) of the lognormal distribution for this study is given by equation (41).

$$f(w_i|\delta_i) = \frac{1}{w_i\sigma\sqrt{2\pi}} e^{-\frac{(\ln w_i - \mu)^2}{2\sigma^2}}, w_i > 0 \qquad (41)$$

where μ = Mean of the shockwaves on the log scale

σ = Standard deviation of the shockwaves on the log scale

Once the marginal pdfs for each lane are found, the optimal Bayes rule to estimate the lane is given by equation (42).

$$\text{if } f(w_i|\delta_i = 1) > \frac{\pi(\delta = 2)}{\pi(\delta = 1)} f(w_i|\delta_i = 2), \quad \delta_i = 1 \text{ else } \delta_i = 2 \qquad (42)$$

where $f(w_i|\delta_i = 1)$ = the lognormal pdf estimated based on training data from SQL

$f(w_i|\delta_i = 2)$ = the lognormal pdf estimated based on training data from LQL

$\pi(\delta = 1)$ = the prior component of group s which is the percentage of s from population $\left(\pi(\delta = 1) = \frac{n(\delta=1)}{n(\delta=1)+n(\delta=2)}\right)$, $n(\delta = 1)$: total number of vehicles observed in SQL, $n(\delta = 2)$: total number of vehicles observed in LQL

$\pi(\delta = 2)$ = the prior component of group s which is the percentage of l from population $\left(\pi(\delta = 2) = \frac{n(\delta=2)}{n(\delta=1)+n(\delta=2)}\right)$

The prior components above were obtained using training data. In the real life situation, training data are a kind of historical data taken from exactly the same location with identical or similar traffic conditions. In this study the training data were generated from the same traffic simulation setting in *Vissim* but with a different seed for random number generation.

3.7.2 Bivariate Mixture Model Clustering

In order to improve the accuracy of the prediction, in addition to the shockwave speeds the distance of the probe to the stop bar is also considered as an input. In that case, a bivariate pdf is needed to characterize the conditional distributions. These pdfs are found by fitting a mixture of normal distributions to the training data which consists of two variables: w_i and d_i which are denoted by vector $x = [w_i, d_i]^T$. Given the lane, the bivariate density is approximated by a mixture of normals as formulated in equation (43),

$$f(x|\delta = j) = \sum_{k=1}^{G_j} \tau_k f_k(x|\delta = j) \quad j \in \{1,2\} \tag{43}$$

where f_k = pdf of mixture component k

j = index for the lane, either 1(SQL) or 2 (LQL).

G_j = number of mixture model components

τ_k = probability that an observation comes from the k-th mixture component. $(\tau_k \in (0,1)$ and $\Sigma_k \tau_k = 1)$

Each component f_k is a bivariate normal distribution [63] and has the probability density function as in equation (44).

$$\phi(x_i, \mu_k, \Sigma_k) = \frac{\exp\left\{-\frac{1}{2}(x_i - \mu_k)^{\mathrm{T}} \Sigma_k^{-1}(x_i - \mu_k)\right\}}{\sqrt{\det(2\pi\,\Sigma_k)}} \tag{44}$$

where μ_k = mean for component k

Σ_k = covariance matrix

In this study, the bivariate mixture model clustering was applied using 2 variables as inputs which are the probe vehicle individual shockwave speeds and the probe vehicle distance in the queue to the traffic light. However, instead of using the mixture model directly for clustering, the model was used to get the best pdf for each lane. These pdfs were then used for the clustering. The clustering was done according to the following steps:

1. Using 2 parameters obtained from the training data (the probe vehicle individual shockwave speeds and the distance to the traffic light while queuing), implement the mixture model to get the best pdf for each lane $\left(f(x|\delta = 1)\ \&\ f(x|\delta = 2)\right)$.
2. Calculate the density for each probe vehicle in the new data using the obtained pdf for each lane.
3. Calculate the prior $\left(\pi(\delta = 1)\ \&\ \pi(\delta = 2)\right)$ as in previous section.
4. Estimate the lane according to the equation (42).
5. Calculate the probability for the prediction confidence. The probability of a probe vehicle to be in the short queue lane is given by equation (45) and likewise for the long queue lane is given by equation (46).

$$P(\delta i = 1) = \frac{\pi(\delta = 1)\, f(x|\delta = 1)}{\pi(\delta = 1)\, f(x|\delta = 1) + \pi(\delta = 2)\, f(x|\delta = 2)} \tag{45}$$

$$P(\delta i = 2) = \frac{\pi(\delta = 2)\, f(x|\delta = 2)}{\pi(\delta = 1)\, f(x|\delta = 1) + \pi(\delta = 2)\, f(x|\delta = 2)} \tag{46}$$

3.8 Results and Discussion

In this section, the results of all the methods discussed earlier are presented. The precisions of each method are calculated to assess the level of accuracy obtained from each model. The precision is computed by the following steps,

1. Categorize the predictions based on the true lane as shown in Table 4.

Table 4. Prediction categories

True lane	Predicted lane	Category
short	short	T1
long	long	T2
short	long	F2
long	short	F1

2. *Calculate the precisions*: The short queue lane identification precisions are calculated using the formula in equation (47) while the long queue lane identification precisions are calculated using equation (48).

$$Pc_s = \frac{T_1}{T_1 + F_1} \tag{47}$$

$$Pc_l = \frac{T_2}{T_2 + F_2} \tag{48}$$

One of the objectives of identifying the probe vehicles' lane in this study is to estimate the queue length properly. In queue length estimation, special attention should be paid to the last probe vehicle. Therefore in this study, the precision is calculated for every last probe vehicles in their predicted lane.

3.8.1 The naïve method

The result of the naïve method for the three scenarios is presented in this section. In scenario 1 (see Figure 10), all the precisions for the LQL are above 90% which is categorized as very good. On the other hand, the best accuracy of the SQL is only 65% (at 90% market penetration),

although the accuracy has surpassed 50% at 20% market penetration rate. The number of cycles exist using probe vehicles are the same for both lanes where at 10%, 20%, and 30% probe vehicles market penetration rate the available cycles are 21, 60, and 84 cycles, respectively. Starting from 40% to 100% market penetration, the available cycles are more than 90.

In scenario 2, the accuracies for LQL prediction are more than 90% for most market penetration except for 20%, 30% and 40% where the accuracies are 87%, 85%, and 89% respectively. This level of accuracy can still be considered as very good. However, as in scenario 1, the result of SQL prediction is not as good as LQL. The best level of accuracies for SQL prediction is reached at 90% market penetration where the precision reaches 73%. Again the accuracies surpass 50% at 20% probe market penetration. The available cycles of SQL and LQL are the same for all market penetration in scenario 2 and they are slightly higher than the ones in scenario 1 although they have the same pattern.

In scenario 3 the level of accuracies for LQL prediction are declining, where the accuracies are about 84 to 85% except for 20%, 40%, and 50% probe vehicle market penetration rate where the accuracies are just 75%, 81%, and 82% respectively. On the other hand, the accuracies of SQL prediction are increasing where all the accuracies are above 50% (reach 60% at 40% market penetration) although the best accuracy is just 67%. As in previous scenarios, the available cycles for both lanes are the same and also higher than previous scenarios.

The result of the naïve method showed that the precisions of the LQL predictions are always higher than the ones of SQL. This is most likely because the overlapping of individual shockwave speed happened more for SQL data distribution region; in other words, the greater the shockwave value, the more likely it will belong to SQL.

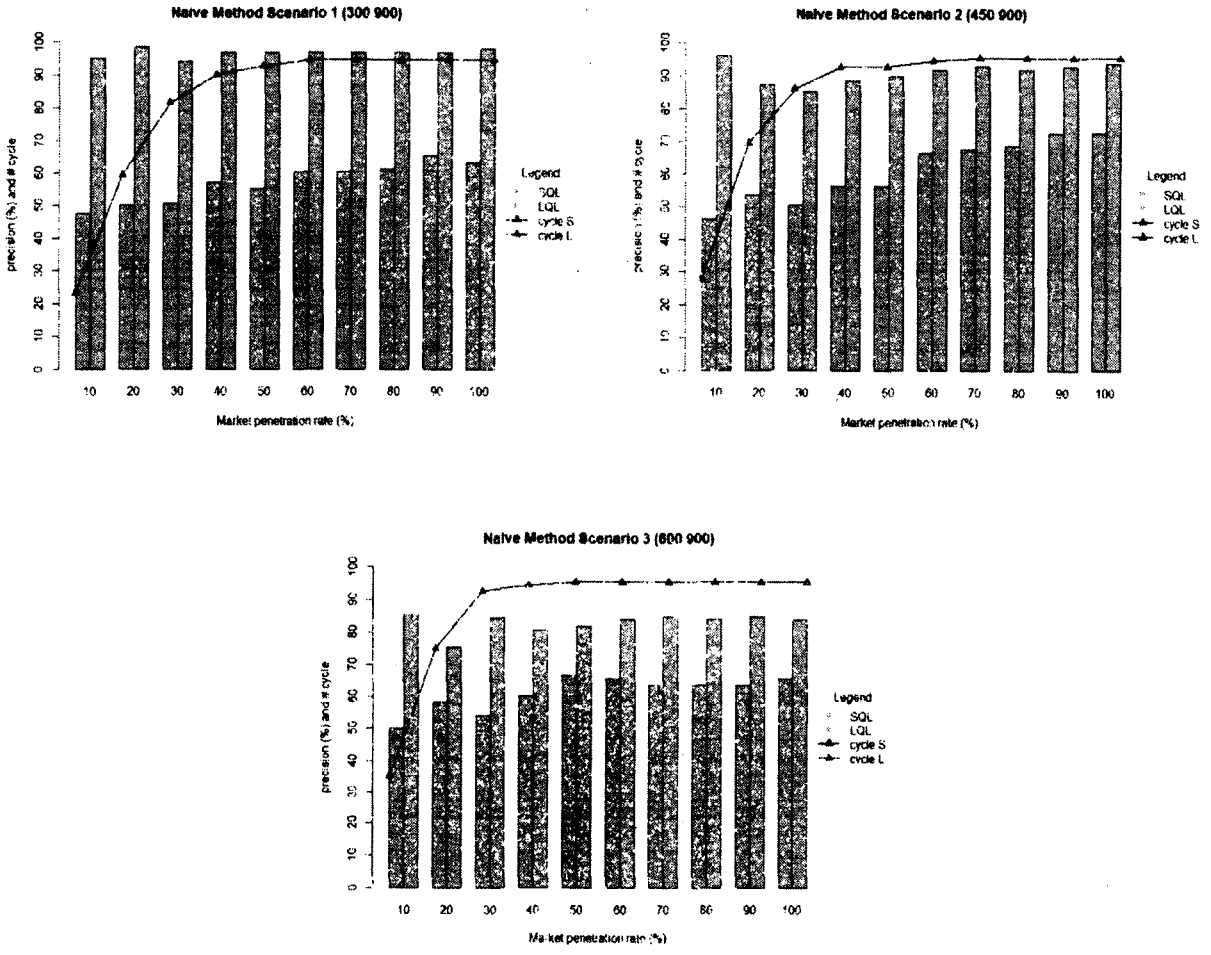


Figure 10. The last probe vehicle prediction precision and available cycle using Naïve Method

Another source of error of using this method is by taking the maximum and minimum individual shockwave values, the naïve method is forced to always have probe vehicle on both lanes. This is a reason to constantly have the same number of available cycles for both lanes in all scenarios, while it might not always be what has really happened, since sometimes there are no probe vehicle available in SQL. The results also show that for scenarios 1 and 2 the precision of the SQL prediction tend to incline as the market penetration rate gets higher. In scenario 3 this tendency can be seen only until 50% market penetration, which most likely is the result of a high number of overlapping between SQL and LQL as the arrival rate of SQL gets closer to the LQL.

The precisions in LQL are relatively stable for all the market penetration rates in all scenarios, which means that the market penetration rates do not have significant effect on LQL lane identification prediction.

3.8.2 The K-Means Clustering Method

The results of K-Means and naïve method clustering show many commonalities. LQL prediction accuracies were higher than those of SQL. This is expected for the reason explained in the previous section. Figure 11 shows the result of the K-means clustering method.

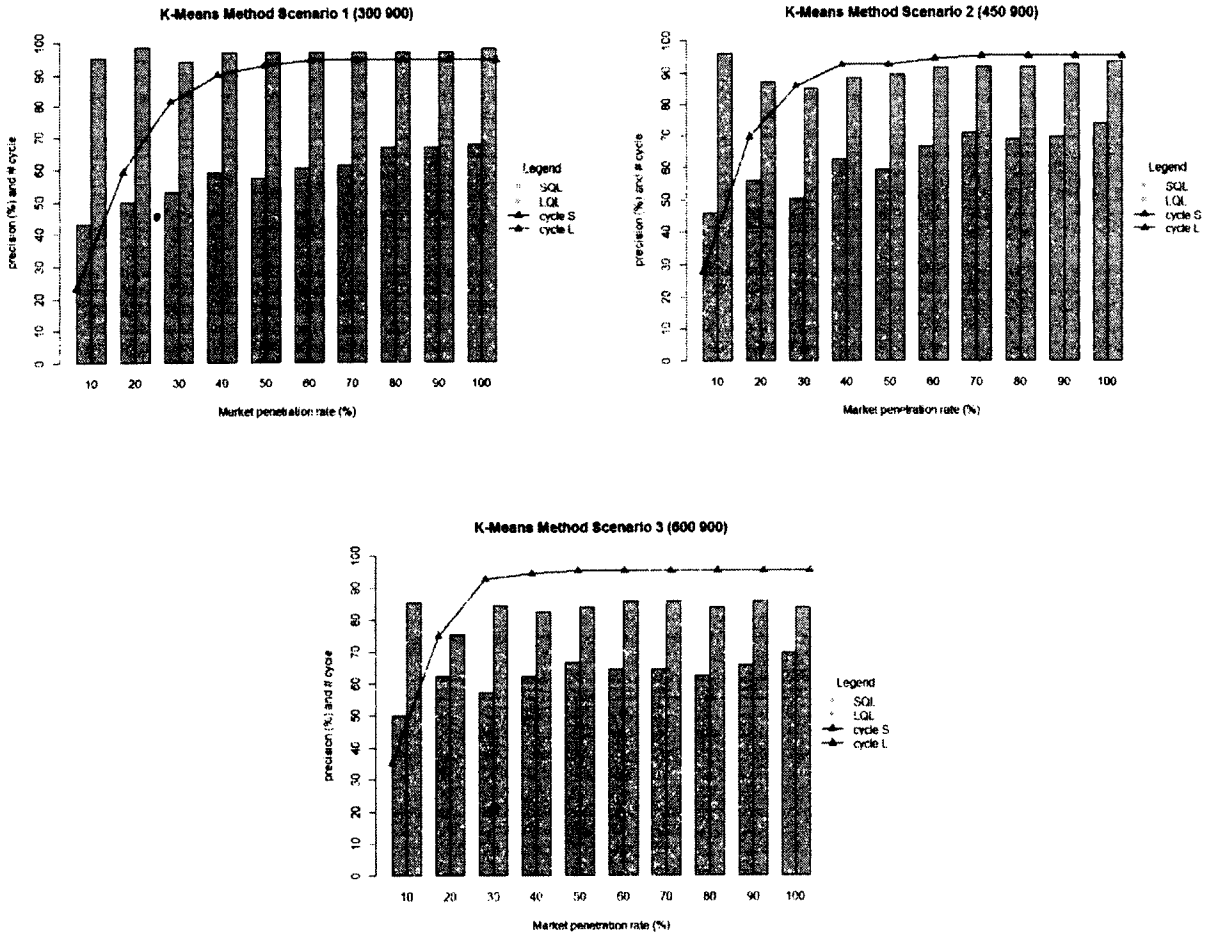


Figure 11. The last probe vehicle prediction precision and available cycle using K-Means Method

In scenario 1, similar to the naïve method, the accuracies of LQL for all market penetration is more than 90%, which is very good. On the other hand, the best accuracies for SQL is 67% (100% market penetration) although the accuracy has been equal to or more than 50% starting from 20% to 100% market penetration rate. For available cycle, both lanes have relatively the same result where starting from 40% market penetration rate there are more than 90 cycles available.

In scenario 2, the accuracies of LQL prediction for almost all market penetration rates are greater than 90% except for 20%, 30%, and 40% market penetration rate where the levels of accuracy are 87%, 85%, and 89%, respectively. The levels of accuracy of SQL prediction surpass 50% at 20% market penetration rate; however, the best level of accuracy is just 74% which is reached at 100% market penetration rate. Again, the available cycles for both lanes are the same, and at the first three market penetration rates, the number of available cycles are slightly higher than the ones in scenario 1. Starting from 40% market penetration rate there are more than 90 cycles available. In scenario 3, the levels of accuracy of LQL prediction are declining to a range of 83% to 86%. At 20% market penetration rate, the levels of accuracy were declining to 75%. On the other hand, the levels of accuracy of SQL prediction in general were increasing. For all market penetration rates, the accuracy level exceeded or equaled 50%, although the highest accuracy level is only 70% (100% market penetration rate). In this scenario almost all market penetration rates have more than 95 available cycles, except for 10% and 20% market penetration rates with 34 and 77 available cycles, respectively.

The similarities of the naïve clustering and K-means clustering results are due to their nature as hard clustering methods, where overlapping clusters is not allowed. Similar to the naïve method, in K-mean the number of the clusters has been determined; this forces the method to always have

a probe vehicle in both lanes. This source of error has become the drawback of both methods, especially in SQL prediction. The effects of market penetration on the lane identification precision using this method are similar to the naïve method, where the effects were more significant to SQL than LQL lane identification.

3.8.3 The Lognormal mixture model

Considering the shortcoming of the previous method as discussed earlier has brought the idea of using soft clustered method. In this study, as explained before, the first soft clustering method is conducted using the lognormal mixture model. The result of this method can be seen in Figure 12.

In scenario 1, as in the previous methods, the accuracies of LQL prediction are greater than 90% for all market penetration rates. On the other hand, the accuracies of SQL for all market penetration are greater than 50%. The least accuracy is 52% (50% market penetration rate) while the best accuracy is 72% (10% market penetration rate). On average, the accuracy of SQL prediction for all market penetration is 63%. Unlike the previous methods, this time the available cycles for each lane is different. The available cycle of the LQL prediction for almost all market penetration are greater than 90 cycles, except for the 10% market penetration which is 75 cycles, while the available cycles for SQL are much less especially in the first three market penetration which are 18, 37, and 52 cycles for 10%, 20%, and 30% market penetration rate, respectively.

In scenario 2, the level of accuracy of LQL and SQL prediction on average were less than the previous scenario. For LQL prediction, the accuracy is 83% for 10% and 20% market penetration, 86% and 89% for 30% and 40% market penetration rate, respectively and the rest are greater than 90%, while for SQL the accuracy for most of the market penetration are above

50% except for 10% market penetration. The highest accuracy level is 71% (100% market penetration rate). For available cycle in LQL, for most market penetration are greater than 90 cycles except for 10% market penetration while for SQL the first two market penetration rate have relatively low available cycle which are 22 and 44 for 10% and 20% market penetration rate.

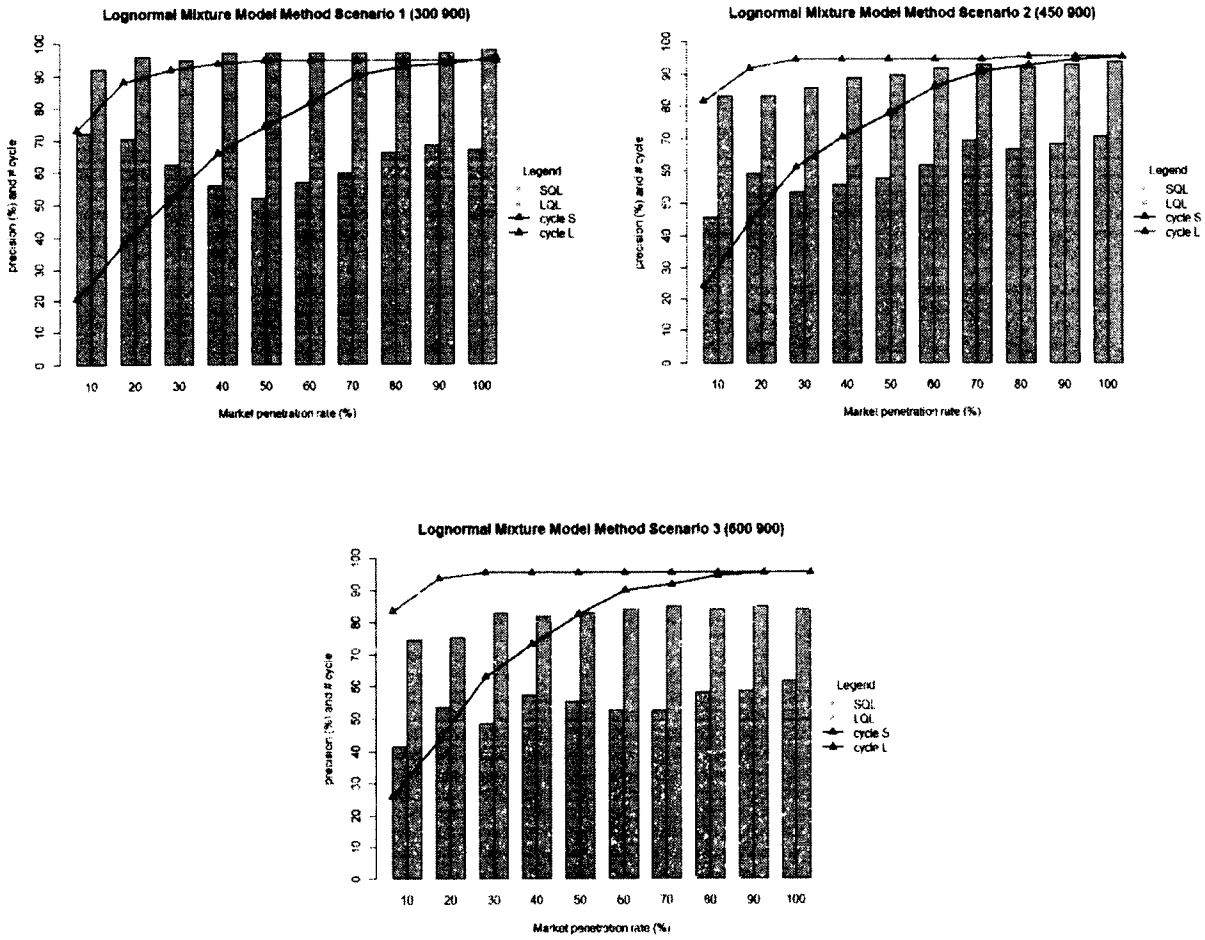


Figure 12. The last probe vehicle prediction precision and available cycle using Lognormal Mixture Model Method

In scenario 3, the precision of LQL as well as SQL predictions decline even more, compared to the previous scenarios. For LQL, the accuracy of the first and the second probe market penetration rates are 74% and 75%, respectively. The rest have accuracies from 82% to 85%. For SQL there are two market penetrations that have accuracies of less than 50% which are the 10% and 30% market penetration, with accuracies 42% and 48%, respectively. The average accuracy for SQL prediction is 54%, while the highest accuracy is 62% for 100% market penetration rate.

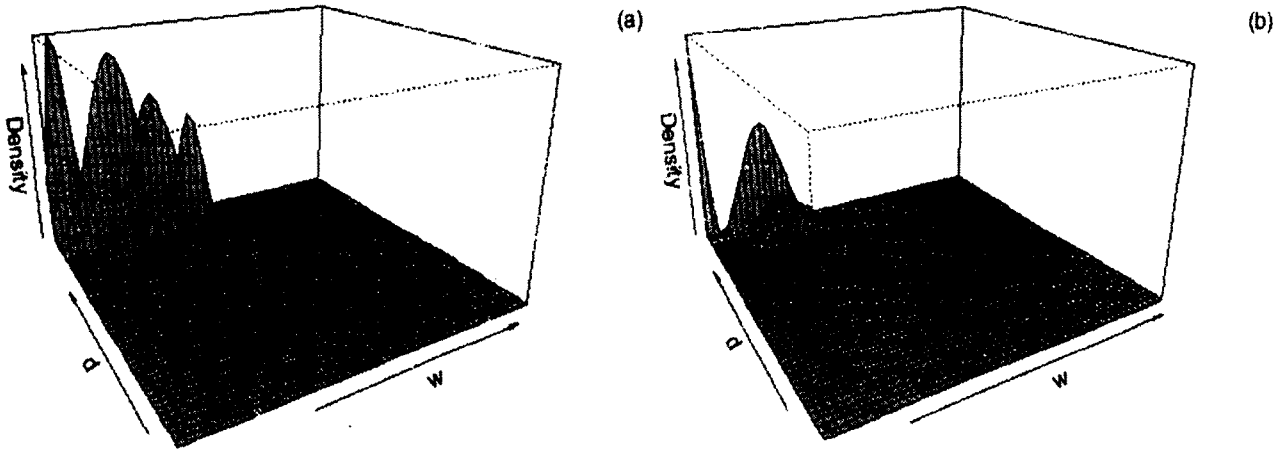


Figure 13. Bivariate mixture model pdf of (a) SQL (b) LQL

The available cycles in this scenario for both LQL and SQL have the same patterns as the ones from previous methods, with slight improvement. Using the lognormal mixture model, although the results are still not as expected, the predictions become more realistic where the cycles available for each market penetration in SQL are not the same as the ones in LQL.

3.8.4 The Bivariate Mixture Model Clustering

Another effort for solving the clustering problem in this study is by conducting bivariate mixture model.

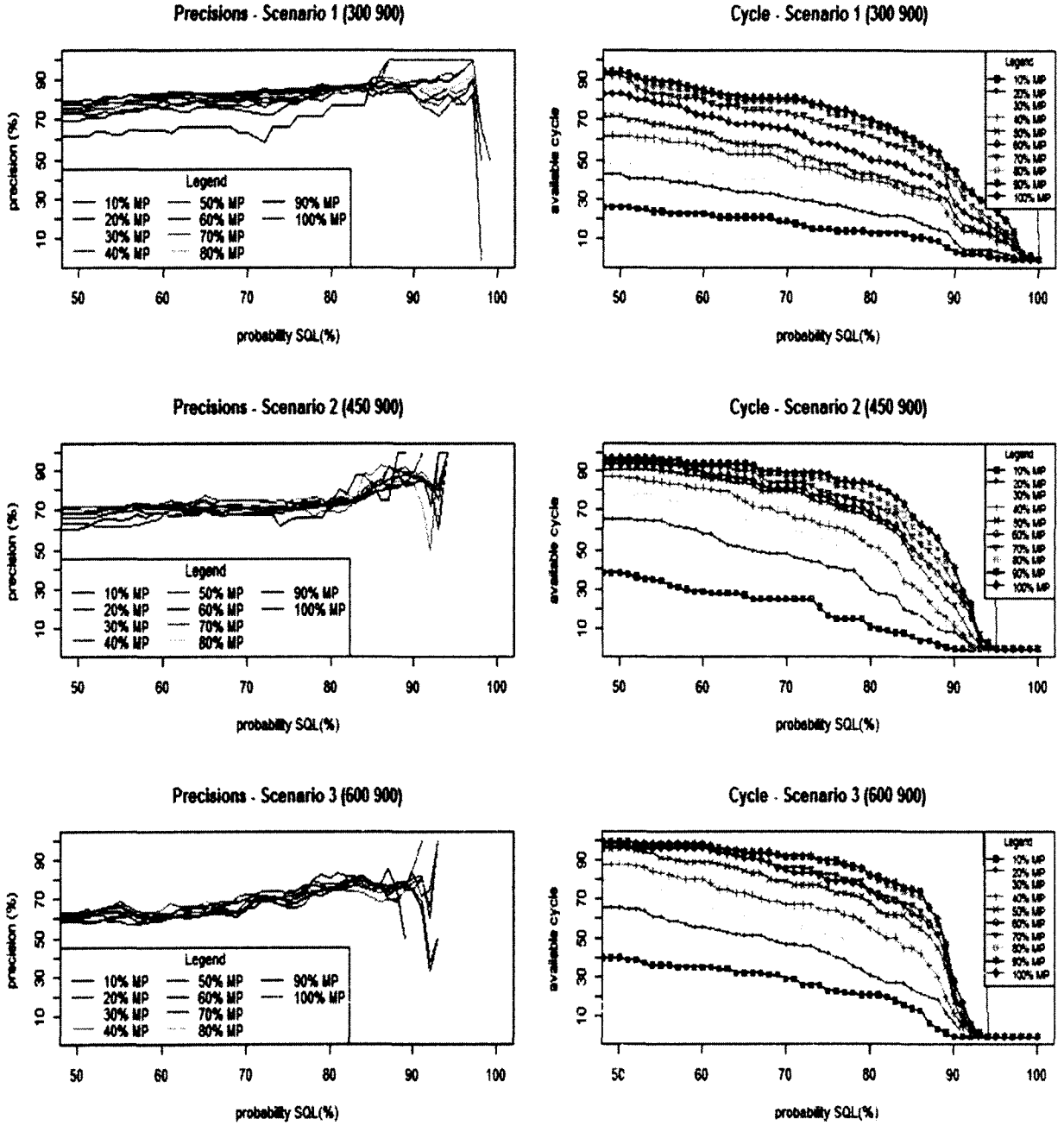


Figure 14. The last probe vehicles' lane prediction precision and available cycle for Short Queue Lane using Bivariate Mixture Model Clustering Method

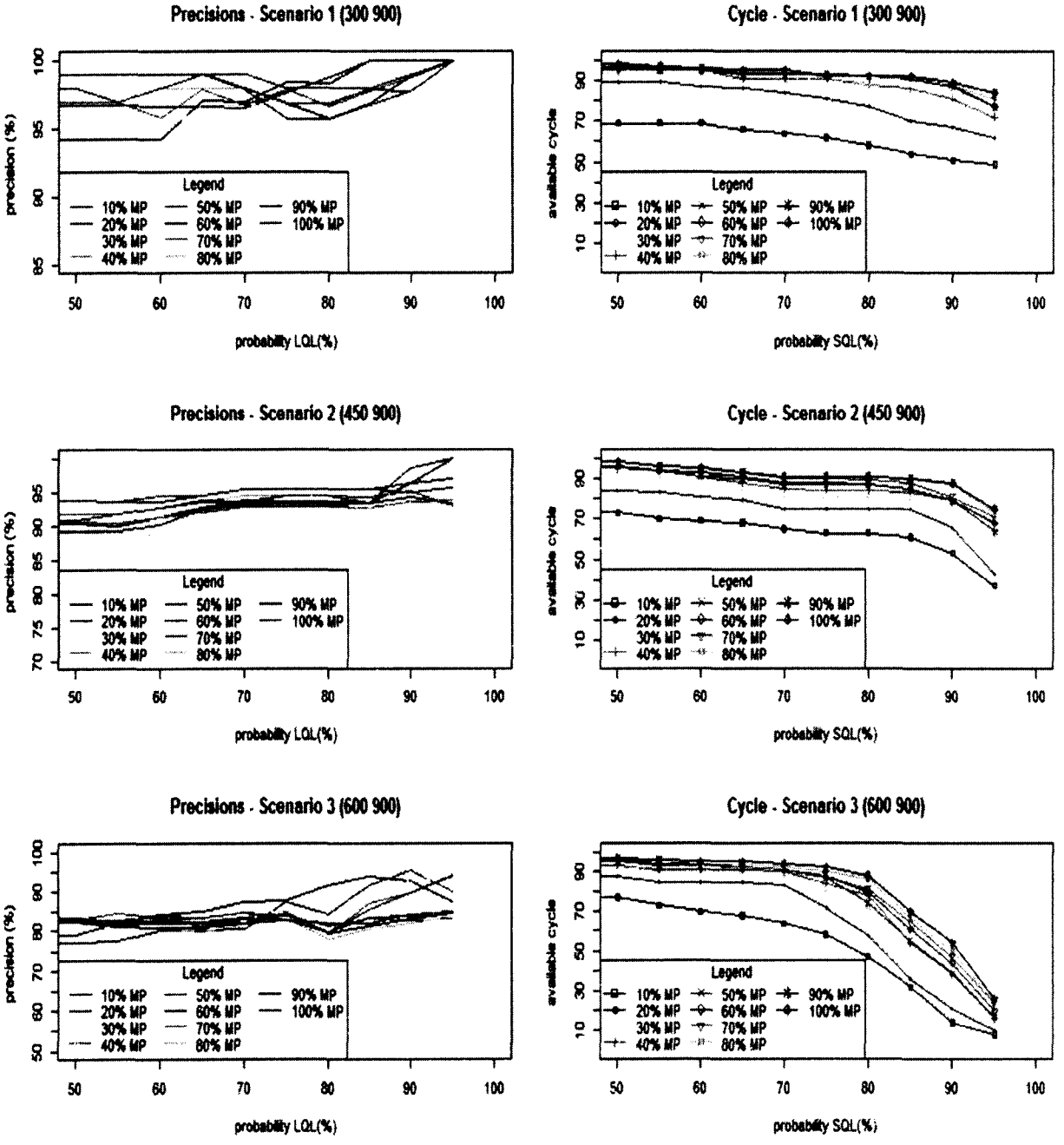


Figure 15. The last probe vehicles' lane prediction precision and available cycle for Long Queue Lane using Bivariate Mixture Model Clustering Method

In this method, beside the individual shockwave speed, the distance was used as another variable. The bivariate mixture model is implemented by utilizing the package mclust [64] in *R* programming software. The model is implemented to get the best pdfs of each lane.

Figure 13 exhibits the example of pdf for each lane estimated using the bivariate mixture model. Figure 13(a) is the pdf for SQL while Figure 13(b) is the pdf for LQL. In this example, based on the shockwave and distance, the SQL data is clustered to become five components while the LQL data become eight components. As explained earlier, the lane identification is not obtained directly from this clustering process. Instead, the model is used only to fit the best distribution for both lanes. Once the best pdfs for both lanes is obtained, the optimal Bayes rule for allocation is implemented for clustering.

In Figure 14, the comprehensive result for the last probe vehicle prediction precisions and available cycles for SQL using Bivariate Mixture Model Clustering Method is presented, while Figure 15 shows the ones for LQL. The horizontal axis (probability) of all the charts in both figures was calculated using the prediction probability (equations (45) and (46)). Both Figure 14 and Figure 15 show that in general for all scenarios, the more the probability the more the accuracy, but with fewer available cycles.

3.8.4.1 Predictability

The probe vehicles' lane identification heavily depends on the individual shockwave speed difference between the two lanes, thus the more the difference, the better the lane prediction accuracy. However, due to the random nature, there were sometimes cycles where it is difficult to distinguish between the individual shockwave of each lane. Figure 16 displays the example of the low prediction probability cycles. The red circles represent the individual shockwaves in SQL, while the black dots are for the ones in LQL. The figure indicates that in those cycles, it is hard to differentiate the probe vehicles' lane position even by taking only the last probe vehicles. The overlapping of the shockwaves makes them undistinguishable. The individual shockwaves in these cycles normally have low prediction probability (as calculated using equations (45) and

(46)) which are usually less than 50%. In other words, using the prediction probability, at a certain threshold which is apparently more than 50%, the identification can be carried out with high confidence.

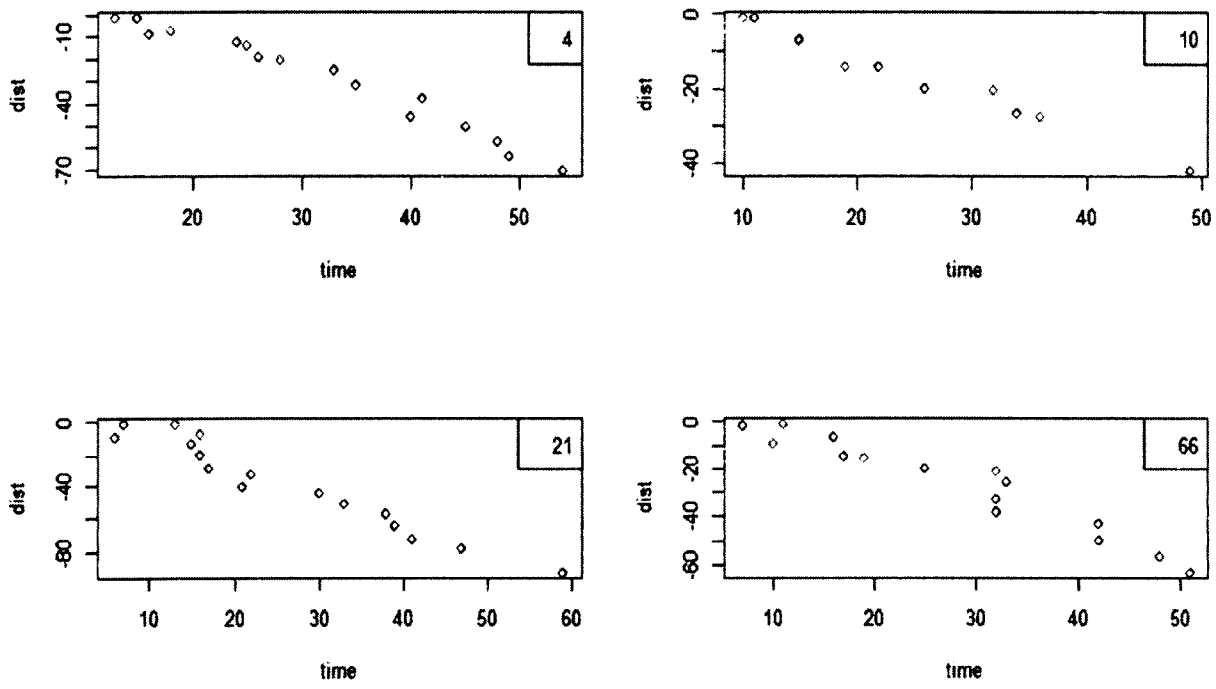


Figure 16. Example of low prediction probability cycles

3.8.4.2 Probability Threshold

As mentioned in the previous section, at a certain degree of probability the probe vehicles' lane prediction can be made with high confidence. However, determining the probability threshold in this study is a bit problematic since there are ten different market penetration rates for three scenarios. Apparently, the desired probability threshold is the probability where the precisions are acceptable but also provide enough number of available cycles for all market penetration

rates. In this study, to determine the probability threshold, the mean precision and also mean available cycle for all market penetration rates were calculated. The result of the calculation is shown in Table 5.

Table 5. The mean of the precision and available cycle of all market penetration in each scenario

Probability Threshold	SQL						LQL					
	Precision			Cycle			Precision			Cycle		
	Sc. 1	Sc. 2	Sc. 3	Sc. 1	Sc. 2	Sc. 3	Sc. 1	Sc. 2	Sc. 3	Sc. 1	Sc. 2	Sc. 3
50	74%	68%	62%	71	84	86	97%	91%	82%	93	93	93
55	76%	69%	64%	67	83	83	97%	91%	82%	93	91	91
60	78%	70%	62%	63	79	82	97%	92%	82%	92	90	90
65	79%	72%	64%	59	76	77	98%	93%	82%	90	88	89
70	79%	71%	68%	58	71	73	97%	94%	83%	90	85	88
75	80%	71%	70%	52	67	69	97%	94%	86%	88	84	83
80	84%	72%	76%	47	59	61	97%	94%	82%	86	84	75
85	87%	82%	76%	41	49	53	98%	94%	86%	84	83	56
90	89%	87%	78%	27	23	17	99%	95%	88%	81	78	39
95	88%	NA	NA	14	0	0	100%	96%	88%	74	63	19
100	NA	NA	NA	0	0	0	NA	NA	NA	NA	NA	NA

Table 5 shows that for SQL at probability 85%, the precision for scenarios 2 and 3 are the best and still acceptable for scenario 1 (>85%). Except for scenario 2, the available cycles for this threshold are greater than 40. For LQL, at the probability 75%, the combination of the precision and the available cycle are pretty good (in all scenarios except 3, the precisions are greater than 85%). Using these considerations, the 85% probability is set as a threshold for SQL, while for LQL the probability threshold set is 75%.

By applying the probability threshold, the prediction precision and available cycle for both lanes are provided by Table 6. This table shows that, on average, the precision of lane identification

for both lanes are good (≥ 85%), except for scenario 3. Most likely this happens since in scenario 3 the arrival for SQL are larger than the other two scenarios, thus this scenario has more overlapping individual shockwaves that make them undistinguishable. Unexpectedly, the available cycles in scenario 2 for both SQL and LQL are the lowest among the three scenarios.

Table 6. The precision and available cycle (threshold 85% SQL and 75% LQL probability)

MP	Precision						Available Cycles					
	SQL			LQL			SQL			LQL		
	Sc. 1	Sc. 2	Sc. 3	Sc. 1	Sc. 2	Sc. 3	Sc. 1	Sc. 2	Sc. 3	Sc. 1	Sc. 2	Sc. 3
0.1	91%	83%	79%	98%	94%	88%	11	6	14	62	63	58
0.2	85%	84%	75%	98%	93%	88%	20	19	24	81	75	72
0.3	92%	83%	73%	97%	93%	85%	24	23	37	88	81	81
0.4	91%	88%	71%	98%	93%	83%	33	32	45	91	84	84
0.5	86%	83%	77%	96%	95%	83%	36	46	57	93	87	87
0.6	82%	80%	75%	97%	95%	83%	44	51	63	93	88	88
0.7	85%	81%	75%	97%	94%	84%	54	57	67	92	90	88
0.8	86%	81%	78%	97%	95%	83%	58	62	72	92	91	90
0.9	87%	78%	79%	97%	95%	85%	61	67	75	92	91	92
1	88%	79%	78%	98%	95%	83%	64	71	76	94	92	94
Average	87%	82%	76%	97%	94%	85%	41	43	53	88	84	83

This means that, in this scenario, there are fewer cycles that have at least one vehicle with prediction probability within the threshold, or in other words, there are fewer cycles that have a probe vehicle whose lane can be identified with high confidence. The result also shows that, except for available cycles, there is no clear relationship pattern between the penetration rate and the lane identification precision.

3.8.5 Lane identification methods comparison

In the previous sections, several methods for the probe vehicles' lane identification have been discussed. The results show that in general,

1. The precision levels of LQL are higher than those of SQL.
2. The lane identification in LQL results in satisfactory precision level.
3. The precision of the lane identification gets higher when the precisions account only for the last probe.

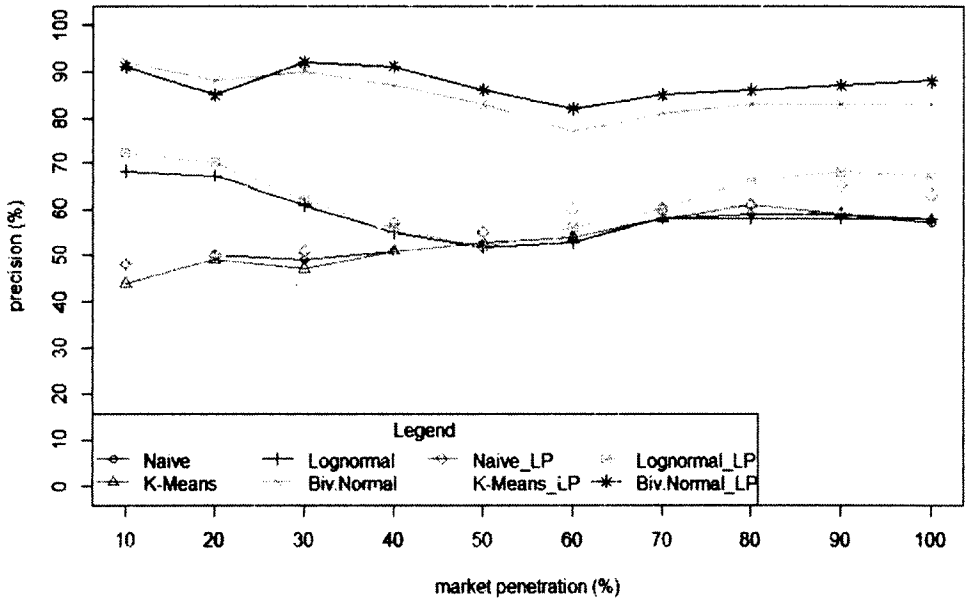


Figure 17. SQL lane identification result comparison (Scenario 1)

Since the SQL lane identification in general has low precision, more attention is given to the prediction for this lane. The result comparison for all methods in scenario 1 is displayed by Figure 17. The comprehensive comparison of all the methods for all scenarios can be seen in Appendix J – Appendix L. The results (also seen in Figure 17) identified that from all the method

discussed previously, the Bivariate normal method is superior compared to other methods. The precision of all the methods above can increased by implementing the confidence level of the results instead of just taking the averages of the precision values. However, one thing that also need to be considered is that filtering the results by taking only the high confidence level will decreased the number of the available cycles. In the Bayes allocation model using the Bivariate mixture model PDFs instead of implementing the confidence level, the high confidence results are obtained from the high probability values.

3.8.6 Summary and conclusion

In this study, a methodology to identify a probe vehicles' lane in a multiple-lane signalized intersection was developed. The probe vehicles' individual shockwave speeds are used to detect their lanes. A naïve method and clustering algorithms were examined to produce the best identification result. Beside the shockwave speed, the probe vehicles' distance was used in a Bivariate Mixture model for clustering their lanes. Instead of clustering data of both lanes in the bivariate mixture model, the model is used to obtain the best probability density function for each lane using both traffic parameters. The probe vehicles' lanes were then identified by implementing the Optimal Bayes Rule for allocation, using the obtained probability density functions.

The result demonstrates that using Bivariate Mixture model, the lane identification can be carried out with acceptable results when the predictability probability is greater than or equal to 85% for short queue lanes and greater than or equal to 75% for long queue lanes. The result also indicates that the more the arrival rate, the less the precision of lane identification.

CHAPTER 4

QUEUE LENGTH ESTIMATION

4.1 Introduction

Queue length is an important traffic parameter needed to optimize the transportation system, especially at the network level. Researchers and traffic engineers have exercised several methods for queue length estimation. The most common method is by using the shockwave speed formed by a vehicles' time and space coordinates when they join the queue. This method typically is applied using a loop detector data as input. However, the installation and maintenance cost of this equipment has prevented the application of these technologies for many arterial networks around the world. Additionally, there is a great chance that the loop detectors will be in error and malfunctioning, as pointed out by [65]. Another drawback of a loop detector is that since most of the time there are only two detectors in a lane of an intersection arm the queuing shockwave line for all the vehicles in between should be projected from these detectors' points. Thus, the shockwave line of a cycle is assumed to be linear, which in reality is not always true.

All these reasons have motivated the transportation engineering community to come up with new technology to collect traffic data which are more reliable and more cost effective. Reliable information can be obtained from vehicle trajectory data which are high-resolution, second-by-second travel data consisting of vehicle time and space information including speed. This information can be produced by the Global Positioning System (GPS) instrumented vehicle system, also known as GPS probe vehicle system. However, this system has its own drawback of this GPS instrument in that it typically yields location information errors. This is a common problem for this equipment. It is well-known that the GPS has accuracy between 3 to 15 meters

[1, 49, 50] and thus instead of forming a neat shockwave line, this system might produce a series of random points that will not shape the shockwave line properly.

Nonetheless, the GPS instrumented vehicle system has been well recognized as an important element in the vehicle communication system for both vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) scenarios because of its capability to provide trajectories of a probe vehicle [65] and its potential for significant reductions in infrastructure cost [66]. As more and more mobile instruments such as smart mobile phone or GPS navigator can be used as probe instruments, more and more proportions of the vehicle population can be employed as probe vehicles. Ideally, if all the vehicles on the road participated as probe vehicles (labeled as 100% probe market penetration), then high-resolution traffic data such as speed, density, flow and even travel time and queue length would become available. These data are needed to optimize the transportation system, for example one would be able to estimate the emissions properly using those data [8].

Alongside its advantages, other than the aforementioned issue, several other items of concern also need to be put addressed. The first is the privacy issue as discussed in [67] and [68]. The GPS instrument provides the time and location information of each probe vehicle and sends these data to a server to be analyzed. Some people consider this as interference with their privacy and thus did not allow their travel information to be known by others. Another issue is, in a large network, there will be vast miles of travel distance and total travel time and also an enormous number of vehicles. This means that for 100% probe vehicle market penetration, there will be a huge amount of high-resolution data collected to the server, which creates other needs such as gigantic storage to accommodate all the data, super computers to carry out the online and offline analysis, and so on.

These issues have changed the perspective that it is not always ideal to have 100% probe vehicle market penetration. The new definition of an ideally-instrumented vehicle application is having a minimum market penetration rate but still being able to optimize the transportation system properly, particularly at the network level. A new challenge for researchers now is how to optimize the use of probe vehicle data resulting from the minimum market penetration rate.

The objective of this study is to determine the optimal probe market penetration rate in the application of dynamic queue length estimation particularly in signalized intersection. The rest of this chapter is organized as follows. Section 2 formulates the problem and states the focus of this chapter. Section 3 provides review of previous studies about queue length estimation. In Section 4, the proposed methodology is introduced including the explanation of the data and simulation, while the results and discussion are described in Section 5. Finally, Section 6 gives the summary and conclusions.

4.2 Problem Definition

Research has been carried out to estimate the queue length, particularly in signalized intersections [56, 61, 69] and metered on-ramps [57, 58]. Most of them were conducted using detector data as inputs. As the probe vehicle is becoming increasingly popular, researchers started to estimate this traffic parameter from probe vehicle data, for example in [5, 53]. However, in these probe vehicles' utilization studies, the approaches were either applied to a single lane, assumed equal queue length for multilane cases, or assumed that the probe vehicles' lane are known, which may not always be true in reality. There are many instances of unequal queues occurring in multiple lanes of a signalized intersection approach, especially when there is a shared lane (i.e. one lane for multiple movements/turns, for example, straight-through, and

right turn lane) in the approach. This study is an attempt to fill this gap, i.e. using probe vehicle data to estimating the queue length for unequal queue situations.

4.3 Related Studies

One of the most common approaches to estimate the queue length is the input-output approach. As indicated by Sharma et al. [70], the maximum queue length can be obtained from the queue polygon by finding the ordinate at the beginning or green plus start-up lost time. Lawson et al. [71] proposed another method to estimate the queue length using the input-output diagram. They modified the input-output diagram to include the number of vehicles that reach the back of the queue in a certain time unit.

However, as stated by Liu et al. [56], the traditional input-output method can only deal with queues that are in the distance between the vehicle detector and intersection stop line since the cumulative vehicle count will not be available at the moment the detector is occupied by the queue. Therefore, instead of using the traditional input-output method, they utilized the queue discharge process and applied the Lighthill and Witham [33] and Richard [34] (LWR) kinematic wave theory to estimate the queue length. Wu and Liu [72] developed a traffic model for a congested arterial network which they called the shockwave profile model using the LWR kinematic wave theory. The method handled homogeneous road segments with constant capacity as a section which differentiates the method from the conventional macroscopic model.

Another approach to estimate the queue length is by using optimization method as demonstrated by [59] where they constructed neural network models to predict intersection queue. Statistical models such as Markov chains also have been employed to model the queue length in signalized intersections as shown in [60, 61]. Beside Markov chains, queue length estimation also has been

conducted using a Kalman Filter such as in [69] where they developed a real-time approach for lane-based queue lengths at isolated signalized intersections while [57] and [58] employed a Kalman Filter to estimate the queue length at a metered on-ramp.

All the aforementioned studies were conducted using loop detector data. Along with the scheme of vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) communication technologies, the use of instrumented (probe) vehicle data became a hot topic in this field. The probe vehicle system is considered to be very useful in dealing with the real time estimation. Researchers began to exercise with probe data in traffic parameter estimation including queue length estimation. One of the first real time queue length estimation studies was conducted by [5]. In their study, they built an analytical formulation based on conditional probability distribution to estimate queue length and its variance. They found that using only the location information of the last probe in a traffic signal cycle, the queue length can be estimated. They also evaluated the model accuracy for each market penetration of probe vehicles. In their later study [54], they tried to quantify the error using this approach. Cetin [15] developed a formula to determine the critical points to characterize the queue dynamic from probe vehicle data and employed the LWR kinematic wave theory to estimate the real time queue length. Ban et al. [53] reconstructed the real time queue length by detecting the critical points in intersection travel times or delay. These studies have somehow proved that the probe vehicle data can be used properly to estimate queue length. However, an essential problem with these studies is that the arrival rate was assumed to be uniform, which does not always happen in reality. An attempt in estimating queue length using probe vehicle without this assumption was done by Ramezani and Geroliminis [17]. In their study, they tried to construct queue profiles from probe vehicle data by clustering them into moving and stopped data and then derived the queue length based on the shockwave profiles.

They claimed that using the estimated queue profiles the arrival distribution patterns and signal settings can be predicted properly. However, the issue of the error measurements of the probe vehicle data was not addressed. As mentioned before, the GPS error may cause the vehicle stop coordinates to look like a bunch of random points rather than a shockwave pattern.

All the aforementioned studies overlooked the lane-based queue length estimation problem by applying the approach to a single lane or by assuming the queue are equal in the case of multiple lanes. This study is an attempt to estimate lane-based queue length using the GPS-equipped probe vehicle in the case of unequal queues by considering the GPS instrument error.

4.4 Methodology

Similar to the previous chapter, the data for this study are generated using transportation simulation software *Vissim* version 6.0. The simulation setting in this research is explained in Section 3.5. As in previous studies in this dissertation, there are three demand scenarios which are 300 and 900 vph, 450 and 900 vph, and 600 and 900 vph for scenarios 1, 2 and 3, respectively.

The LWR shockwave theory is employed to estimate the queue length using the probe vehicle information. The application of the theory for queue length estimation is discussed in detail in Section 2.6. The probe vehicle time and space coordinates of the probe vehicle when they join the queue and also when they discharge were used to estimate the queue length.

Considering the GPS instrument error, ahead of estimating the queue length, the probe vehicles' lanes need to be identified. The probe vehicles' lane identification process is discussed comprehensively in chapter 3.

In this research, the lane identification is conducted by adopting the Bivariate Mixture Model Clustering as describe in Section 3.7.2. Following lane identification as discussed in previous sections, the queue length estimation is conducted using the probe information of each lane which is the one with short queue length (labeled as short queue lane / SQL) and the one with long queue length (labeled as long queue lane / LQL). The limitation of LWR shockwave theory is that the queue length was defined by the intersection between queuing shockwave and discharging shockwave. This means this theory can only predict the maximum queue that can take place instead of the real queue length.

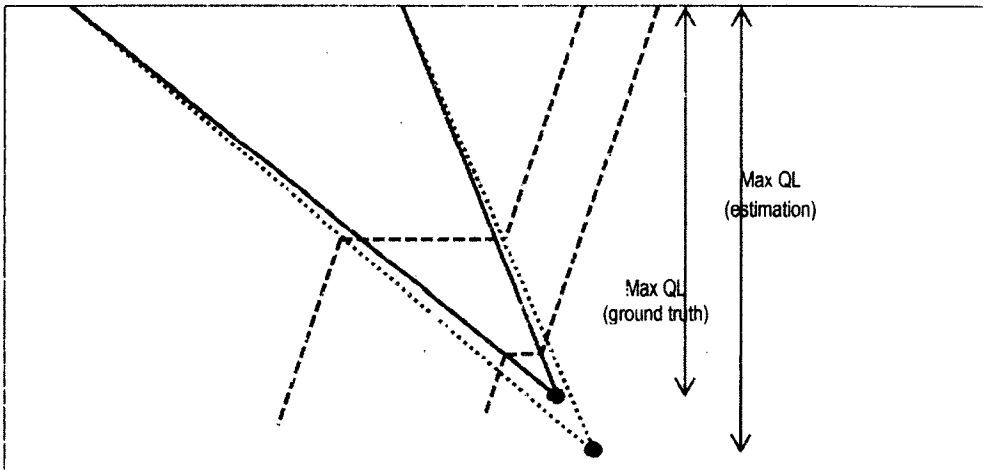


Figure 18. Maximum queue length estimation using LWR shockwave theory

Considering this limitation, the error of the queue length estimation is obtained by comparing the maximum queue length calculated from the true last probe vehicle and the estimated one, as illustrated in Figure 18. The queue length estimation error is obtained by comparing the vertical distance of the shockwave intersection points as shown in the figure.

4.5 Result and Discussion

Mean absolute error (MAE) is utilized to measure the bias between the estimation and the ground truth. The MAE are calculated according the formula in equation (49).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{49}$$

where n = number of samples

e_i = error of sample i

The MAE values are compared to the shockwave mean value to understand the proportion of error (labeled as mean absolute percentage error/MAPE). In this research, the probe market penetration is varied from 10% to 100% by increments of 10%. The more the market penetration, the more the number of available cycles expected. The lane-based MAE and MAPE values and also the number of available cycles of each market penetration for each scenario are presented in Figure 19. As expected, the figure shows that in all cycles for all market penetration the number of cycles for SQL are always less than the number of cycles for LQL. The figure also exhibits that the MAE for LQL is less than SQL for all scenarios where the MAPE for LQL for scenarios 1 and 2 reach below 10% in 40% market penetration and for scenario 3 in 70% market penetration. Although the MAPE of SQL are a bit large, but actually the bias is tolerable except for scenario 3, where on average the bias for scenarios 1, 2, and 3 are 5 meters, 16.4 meters, and 23.1 meters, respectively. Appendix N provides full details about the result.

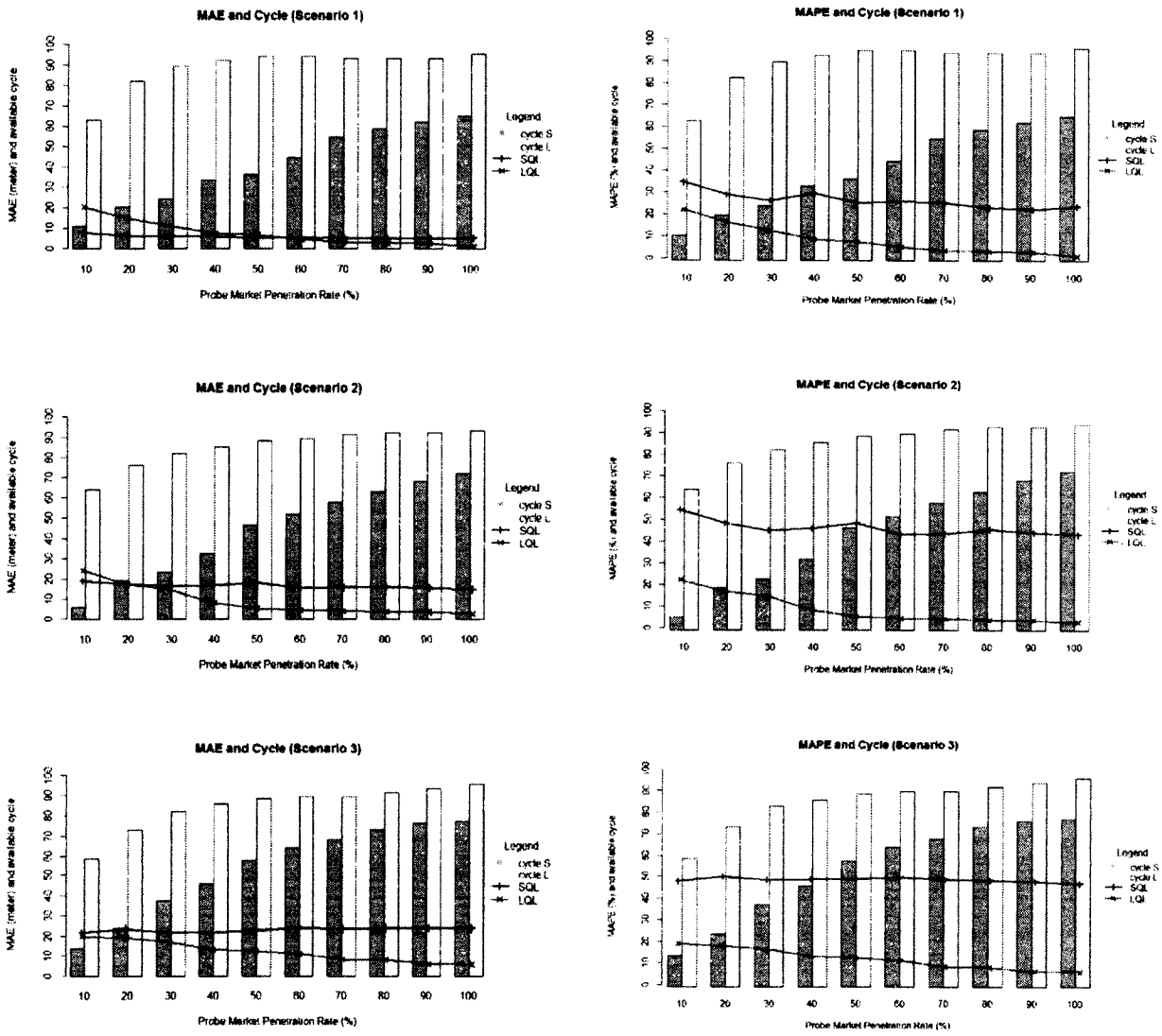


Figure 19. Queue length estimation error and available cycles

One reason that the SQL results are not very good is that in the lane identification process, the vehicles that are taken to be identified properly are the vehicles that comply with a probability threshold of 85% for SQL and 75% for LQL, while all the others that do not meet these thresholds were filtered out. Because of this process, the last vehicle identified in a lane might not really be the last vehicle estimated in the cycle, which produces more bias.

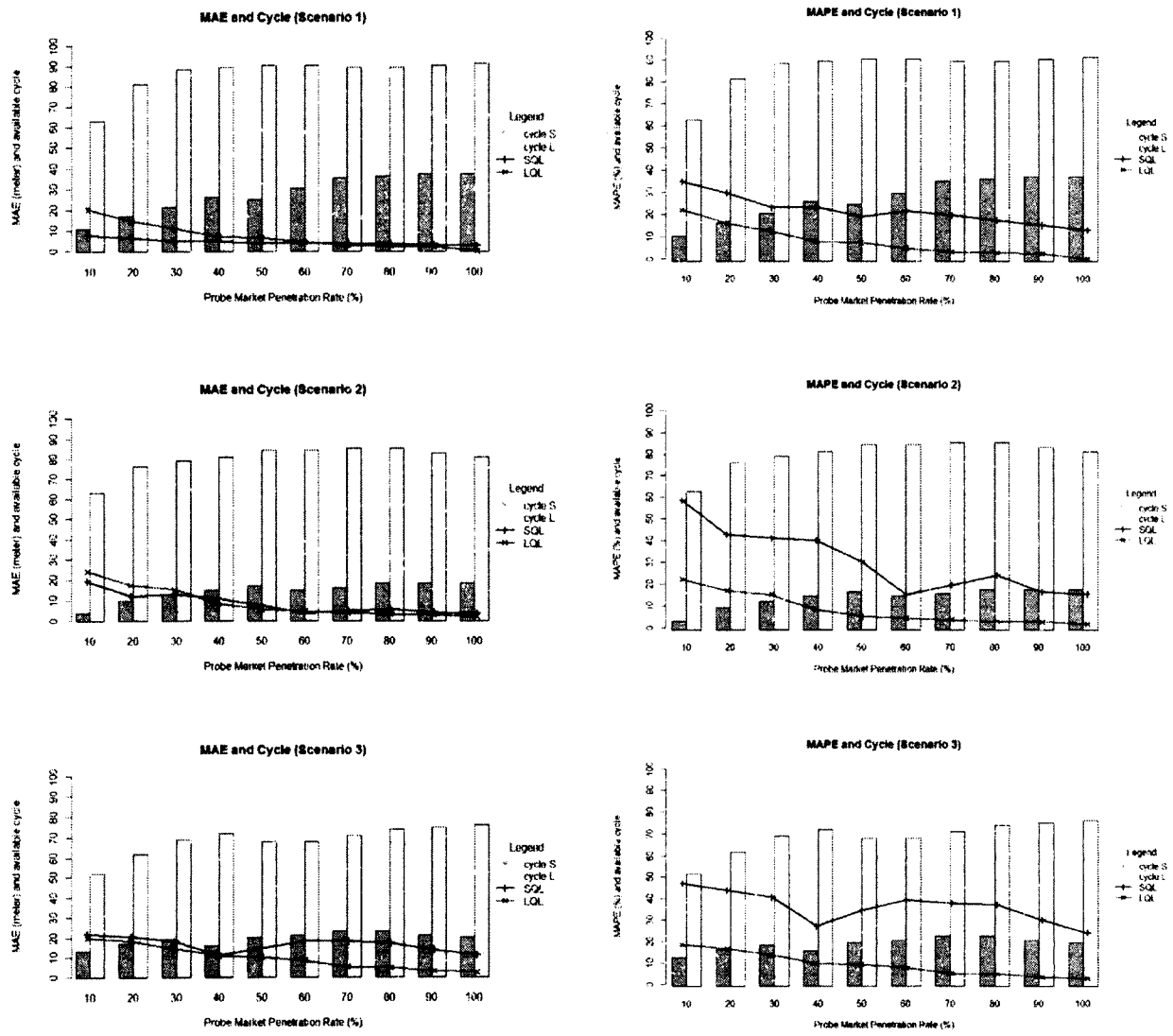


Figure 20. Queue length estimation error and available cycles by determining the last probe first

To deal with the aforementioned issue, the process order is reversed, i.e. the last probe vehicles are determined first and then the probability threshold is applied after that. This is done to decrease the MAE, however it is worth noticing that there is tradeoff between the improvement of accuracy and the cycle availability as shown in Figure 20. The figure confirms the improvement in estimation accuracies, where the biases for SQL queue length estimation on

average are 4.5 meters, 8.3 meters, and 16.4 meters for scenarios 1, 2 and 3, respectively. Meanwhile, for LQL, the MAPE value are equal to or below 10% starting from 40% market penetration for all scenarios. However, as stated above, the number of cycles decreases significantly for SQL for almost all market penetration rates in all scenarios. The comprehensive results for this method are provided by Appendix Q.

Another important finding to mention is that, using all approaches, the queue length estimation accuracy of LQL is always better when compared to SQL. The best explanation for this is that the queue length estimation precision depends much on the accuracy of the lane identification. As has been demonstrated by the result in the lane identification chapter, the accuracies of probe vehicles' lane identification for LQL are better than SQL for all scenarios. The results show that the more SQL demand, the more inaccurate the queue length prediction. Again, this tends to have a linear relationship with the lane identification, where the more the demand the less accurate the lane identification. The explanation of this is the higher the demand for SQL the more chance that the shockwave speeds between both lanes overlap which lessens the probability for their lane to be identified correctly.

4.6 Summary and Conclusion

This study has demonstrated the uses of probe vehicles in estimating queue length. Ten probe market penetration scenarios in three demand scenarios have been built to generate data using the traffic simulation software. The LWR shockwave theory has been utilized successfully to estimate the queue length following the probe vehicles' lane identification.

The result shows that in most cases using 40% market penetration rate is enough to reach about 90% accuracy. The results also confirm that lane identification plays an important role in

estimating queue lengths of unequal queues, which can be predicted well if the lane identification is conducted properly.

CHAPTER 5

FUEL CONSUMPTION ESTIMATION

5.1 Introduction

According to the US Environmental Protection Agency (EPA), in 2011 the transportation sector contributed approximately 28% of the total U.S. greenhouse gas emissions which is the second largest emissions contributor after the electricity sector. Furthermore, as indicated by [73] vehicle emissions is the main cause of nearly 53,000 early deaths in the U.S. annually. These facts have made vehicle emissions and fuel consumption a critical issue in the area of transportation engineering. Thus, there is an obvious need for finding the methodology in estimating the fuel consumption and emissions properly.

Related to this, researchers have built emissions model such as MOBILE [74, 75] which later was replaced by MOVES [44]. The latter model is capable to model the emissions at the microscopic level. The traffic parameters, i.e. speed and acceleration, are the main factors that impact the level of fuel consumption and emissions in this modeling level.

Probe vehicle system is an appropriate Intelligent Transportation Systems (ITS) application to collect high resolution (second-by-second) traffic parameter data. This system is capable of providing the probe vehicles' speed as well as the time and space information known as vehicles trajectories. Using these data, the instantaneous speed and acceleration level can be captured. Currently, researchers have used the probe vehicle to estimate the travel time, where relatively accurate travel estimation can be achieved using low probe vehicle market penetration rate. Due to more and more mobile instruments being used as GPS instruments, higher penetration rates

can be achieved. With a high enough penetration rate, researchers have been able to estimate the queue length using probe vehicle trajectories information. However, the number of studies conducted for implementing the probe vehicle in fuel consumption and emissions estimation is still very limited. Furthermore, those studies were still not able to predict lane-based emissions estimation. The capability to carry out the lane-based emissions estimation is essential because there are a lot of real-life situations where the queue lengths are not the same in the case of multiple lanes. This study is an attempt to estimate the fuel consumption in the case of unequal queue at multilane signalized intersection using the probe vehicle information. This chapter was organized as follows: the next section formulates the issue that will be solved in this study, in Section 3, related studies are discussed, Section 4 provides the methodology, while Section 5 presents the results and discussion. Finally, Section 6 summarizes and concludes the study.

5.2 Problem Definition

This study is an attempt to properly estimate the fuel consumption using probe vehicles by considering the position of the vehicle in the lane while it was queuing. Benefitting from this knowledge, fuel consumption can be estimated properly, especially in unequal queue length cases. Using the fuel consumption by knowing a constant, the CO_2 emissions can be calculated since the emissions is related proportionally to the fuel consumption. This study also is an attempt to have a good description on the probe vehicle market penetration amount that is considered as adequate to predict the emissions accurately.

The question that is tried to be answered in this study are formulated as follows:

- How significant is the role of probe vehicles' lane identification in fuel consumption estimation?

- What is the sufficient probe market penetration rate to predict the fuel consumption properly?

5.3 Related Studies

Several efforts have been made to estimate the vehicle emissions and fuel consumption, for example the implementation of statistical model to estimate the emissions as demonstrated by [76]. Related to this, researchers are also developing emissions model such as MOBILE [74, 75] which is a macroscopic emissions model sponsored by the EPA. Later, this model was replaced by MOVES (motor vehicle emission simulator) which is an emissions model that can be used at the macroscopic, mesoscopic, and microscopic levels, as indicated by [45]. Studies also have been conducted to increase the accuracy of fuel consumption and emissions model in macroscopic emissions model by replacing the speed average value by mean speed distribution, for example as shown in a study conducted by [77]. Another study carried out in Europe [78] about the sensitivity of the pollutant estimation to the driving speed confirmed this idea and claimed that the precision and quality of driving speeds description will impact the pollutant emissions estimation significantly. Furthermore, the authors claimed that applying speed distribution instead of the average values into the emissions functions will result in fuel consumption and emissions estimations that are 14% lower in urban areas, 30% and 20% higher, on rural roads and on motorways, respectively.

To increase the pollutant estimation accuracy, [31] estimates the vehicle and fuel consumption based on instantaneous speed and acceleration levels. The accuracy gets higher since the latter technique was done microscopically or in more detailed level. In this fuel consumption estimation, the speed data involved were not average speeds or mean speed distribution as usually applied in a macroscopic way but second-by-second speed data that can be used to

capture the acceleration and deceleration. These data are one of the essential factors to predict vehicle fuel consumption and emissions properly. Using this model, in his Ph.D. dissertation, [36] developed the mesoscopic model for fuel consumption and emissions estimation. However, in this model, the input was built by constructing a synthetic drive cycle and then estimated the proportion of the average time that a vehicle spends cruising, decelerating, stopping (idling) and accelerating while traveling on a link.

5.4 Methodology

5.4.1 Data

Similar to the vehicle identification study, the data for this part of the research are generated using transportation simulation software called *Vissim* version 6.0. The simulation setting in this research is explained in Section 3.5. As in previous studies in this dissertation, there are three demand scenarios which are 300 and 900 vph, 450 and 900 vph, and 600 and 900 vph for scenarios 1, 2, and 3, respectively.

The LWR shockwave theory is employed to estimate the queue length using the probe vehicle information. The application of the theory for queue length estimation is discussed in detail in Section 2.6. The probe vehicle time and space coordinates of the probe vehicle when they join the queue and also when they discharge from the queue were used to estimate the queue length.

5.4.2 Lane identification

Considering the GPS instrument error, ahead of estimating the queue length, the probe vehicles' lane needs to be identified. The probe vehicles' lane identification process is discussed comprehensively in chapter 3. In this research, lane identification is conducted by adopting Bivariate Mixture Model Clustering, as described in Section 3.7.2.

5.4.3 Synthetic trajectories reconstruction

As trajectories of all vehicles in the intersection are needed to estimate the fuel consumption, the subsequent step after the lane identification is the construction of synthetic trajectories, which are trajectories that are constructed based on the probe vehicles' information. The synthetic trajectories are constructed according to the following steps:

1. Using the probe vehicle information when they are starting joining and discharging from the queue, create the shockwave profile.
2. Estimate the number of vehicles whose trajectories are needed to be developed. There will be three condition types to reconstruct the synthetic trajectories. The first is the ones between the traffic light and the first probe which is estimated by equation (50),

$$n_f^c = \frac{d_1^c}{s} \tag{50}$$

where n_f = number of vehicles in between the traffic light and the first probe vehicle at cycle c.

d_1^c = distance of the first probe to the traffic light at cycle c

s = spacing headway

The second is the ones between two probe vehicles, estimated by equation (51),

$$n_{i,i+1}^c = \frac{d_{i,i+1}^c - s}{s} \tag{51}$$

where $n_{i,i+1}^c$ = number of vehicles between probe vehicles i and $i+1$ at cycle c

$d_{i,i+1}^c$ = distance between probes vehicle i and $i+1$ in cycle c

s = spacing headway

The last is the ones between the last probes and the maximum queue, estimated by equation (52),

$$n^c_{i,q_{max}} = \frac{d^c_{i,q_{max}} - s}{s} \tag{52}$$

where $n^c_{i,q_{max}}$ = number of vehicles in between probe vehicles i and q_{max} at cycle c

$d^c_{i,q_{max}}$ = distance between probe vehicles i and $i+1$ in cycle c

s = spacing headway

3. Calculate each synthetic vehicle's distance, when it joins the queue, to the traffic light. If there is any vehicle between the first probe vehicle and the traffic light, the first synthetic vehicle is considered to have zero (0) distance with the traffic light. Meanwhile, the distance for the rest of the synthetic vehicles (if any) in this gap is calculated as in equation (53).

$$ds^c_j = \frac{d^c_1}{n^c_f} \times \left(id^c_j - 1\right) \tag{53}$$

where ds^c_j = distance of synthetic vehicle i to the traffic light at cycle c

d^c_1 = distance between the first probe vehicle at cycle c and the traffic light

id^c_j = index of the synthetic vehicle j at cycle c

If the gap between probe vehicles is not the last gap, the distance of a synthetic vehicle to the traffic light is calculated as equation (54)

$$ds_j^c = \frac{d_{i,i+1}^c}{n_{i,i+1}^c + 1} \times id_j + d_i^c \quad (54)$$

where d_i^c = distance of nearest (in the gap) probe vehicle i to the traffic light at cycle c

In case of the last gap, the distance of a synthetic vehicle, when it started joining the queue, to the traffic light is calculated as in equation (55),

$$ds_j^c = \frac{d_{i,q_{max}}^c}{n_{i,q_{max}}^c + 1} \times id_j + d_i^c \quad (55)$$

4. Calculate each synthetic vehicle stop time. If the synthetic vehicle is not in the last gap, the stop time of synthetic vehicle j at cycle c is calculated as in equation (56),

$$st_j^c = \frac{lt_{i,i+1}^c}{n_{i,i+1}^c + 1} \times id_j + st_i^c \quad (56)$$

where st_j^c = stop time of a synthetic vehicle j at cycle c.

$lt_{i,i+1}^c$ = lapse of time between stop time of probe vehicle i and $i+1$ atcycle c

st_i^c = stop time of probe vehicle i at cycle c

In case the synthetic vehicle is in the last gap, its stop time is calculated as in equation (57),

$$st_j^c = \frac{lt_{q_{max},i}^c}{n_{i,i+1}^c + 1} \times id_j + st_i^c \quad (57)$$

where $lt_{q_{max},i}^c$ = lapse of time when queue reaches maximum queue length and stop time of probe vehicle i. The time when the queue reaches maximum can be obtained by equation (58),

$$st_{max}^c = \frac{QL_{max}^c}{ws_{lp}^c} \quad (58)$$

st_{max}^c is the time associated with maximum queue at cycle c (QL_{max}^c) calculated using LWR model and ws_{lp}^c is the associated queuing shockwave (calculated using the last probe).

5. Calculate the discharging time for each synthetic vehicle. This information is calculated using the same procedure except that obviously this procedure will utilize the probe vehicle discharging time instead of the stopping time.
6. Once all this information is obtained, build the synthetic trajectories of each synthetic vehicle. There are four types of movements in these synthetic trajectories: cruising, deceleration, stopping and acceleration. In this case, the cruising itself is divided into two parts: cruising before stopping and the cruising after stopping. The total cruising duration was obtain by equation (59),

$$cr = \frac{d_l - d_{acc} - d_{dec}}{v} \quad (59)$$

where cr = total cruising duration for each synthetic vehicle (s)

d_l = link length

d_{acc} = acceleration distance

d_{dec} = deceleration distance

v := average speed (m/s)

The total travel time for each synthetic vehicle is calculated according to the equation (60),

$$ttt_j = cr_j + dr + str_j + ar \tag{60}$$

where ttt_j = total travel time of synthetic vehicle j

dr = deceleration duration

str_j = stop duration; discharge time – stop time of synthetic vehicle j $\left(dt_j^c - st_j^c\right)$

ar = acceleration duration

The cruising duration before stopping is calculating by equation (61),

$$cr_{bs} = \frac{\left(sg - ds_j^c\right) - d_{dec}}{v} \tag{61}$$

where cr_{bs} = cruising duration before stopping (s)

sg = signal distance from upstream

ds_j^c = distance of synthetic vehicle j to the signal when it joins the queue.

5.4.4 Acceleration and deceleration modeling

The input for estimating the fuel consumption is the spontaneous speed, therefore after the synthetic trajectories reconstruction, the synthetic speed profile for the unobserved vehicle needs to be modeled. The idea of the synthetic speed profile is adopted from the study conducted by [36]. There are four types of speed for the synthetic vehicle, which are the cruising, decelerating, idling/stopping and accelerating. For a stopping vehicle, the speed = 0 km/h while for cruising vehicle the speed = free flow speed. Thus, the acceleration and deceleration speed profiles are the only ones that need to be modeled. Using training data generated from exactly the same simulation setting with different seed, the acceleration and deceleration speed profiles were extracted. The data extraction for deceleration speed profile was carried out through the following algorithm:

1. Subset data for a vehicle (using training data) that experiences stopping.
2. Find the coordinates (time and space) when it is stopped (speed = 0).
3. Find the cruise coordinates before vehicle stopped (speed > (Average Speed – Speed standard deviation) & time stamp < time stamp when it stops).
4. Subset the deceleration data which is the data between last cruise coordinate and first stop coordinate.
5. To make sure that data are taken from the same stop (vehicle might have stopped twice), subset the deceleration data only for the ones before the first stop.

While the data extraction for acceleration speed profile are carried out through the following algorithm:

1. Subset data for a vehicle (using training data) that experiences stopping.

2. Find the coordinates (time and space) when it stopped (speed = 0).

3. Find the cruise coordinates *after* vehicle stopped (speed > (Average Speed – Speed standard deviation) & time stamp > time stamp when it stops).

4. Subset the acceleration data which is the data between last stop coordinate and first cruise coordinate (after stop).

5. To make sure that data taken are from the same stop (vehicle might have stopped twice) the acceleration data was subset only for the ones with space coordinate after the stopping space coordinate and before the second stop (if there is any).

The extracted deceleration and acceleration data are used to model the deceleration and acceleration. Both are modeled using linear, quadratic, logarithmic, exponential, power, and polynomial regression. The criterion for the chosen model is the model with the highest coefficient of determination (R^2) value. The next information needed is the acceleration duration and deceleration duration, both obtained by taking the mean of the associated data.

5.4.5 Fuel Consumption Estimation

The Virginia Tech Comprehensive Power-Based Fuel Consumption model (VT-CPFM) model is utilized to estimate the fuel consumption. This model was discussed in detail in [42, 79]. The fuel consumption VT-CPFM model is formulated as in equations (62) - (64),

$$R(t) = \frac{\rho}{25.92} C_D C_h A_f v(t)^2 + 9.8066m\frac{C_r}{1000}(c_1 v(t) + c_2) + 9.8066mG(t) \qquad (62)$$

$$P(t) = \left(\frac{R(t) + 1.04\,\mathrm{ma}(t)}{3600\eta_d}\right) v(t) \tag{63}$$

$$FC(t) = \begin{matrix} \alpha_0 + \alpha_1 P(t) + \alpha_2 P(t)^2 & \forall P(t) \geq 0 \\ \alpha_0 & \forall P(t) \leq 0 \end{matrix} \tag{64}$$

where $R(t)$ = resistance force (N)

ρ = the density of air at sea level at 15°C = 1.2256 kg/m^3

C_D = vehicle drag coefficient (unitless)

C_h = correction factor for altitude (unitless) = 1-0.085 H; H is the altitude (km)

A_f = vehicle frontal area (m^2)

C_r, C_1, C_2 = rolling assistance parameters

m = vehicle mass (kg)

$a(t)$ = vehicle acceleration (m/s^2) at time t

η_d = driveline efficiency

G = roadway grade

In this study, the roadway grade is set to 0 (flat road) and the vehicles are assumed to be uniform and the type is Toyota Camry. Both probe vehicles' and synthetic vehicles' speed information were used as inputs to calculate fuel consumption.

5.5 Results and Discussion

5.5.1 Synthetic vehicle trajectories construction

Basically, the synthetic trajectories were formed by assuming that both queuing and discharging shockwave lines in between probe vehicles are linear. Likewise for the lines between the traffic

light and the first probe and between the last probe and calculated maximum queue as shown by Figure 21.

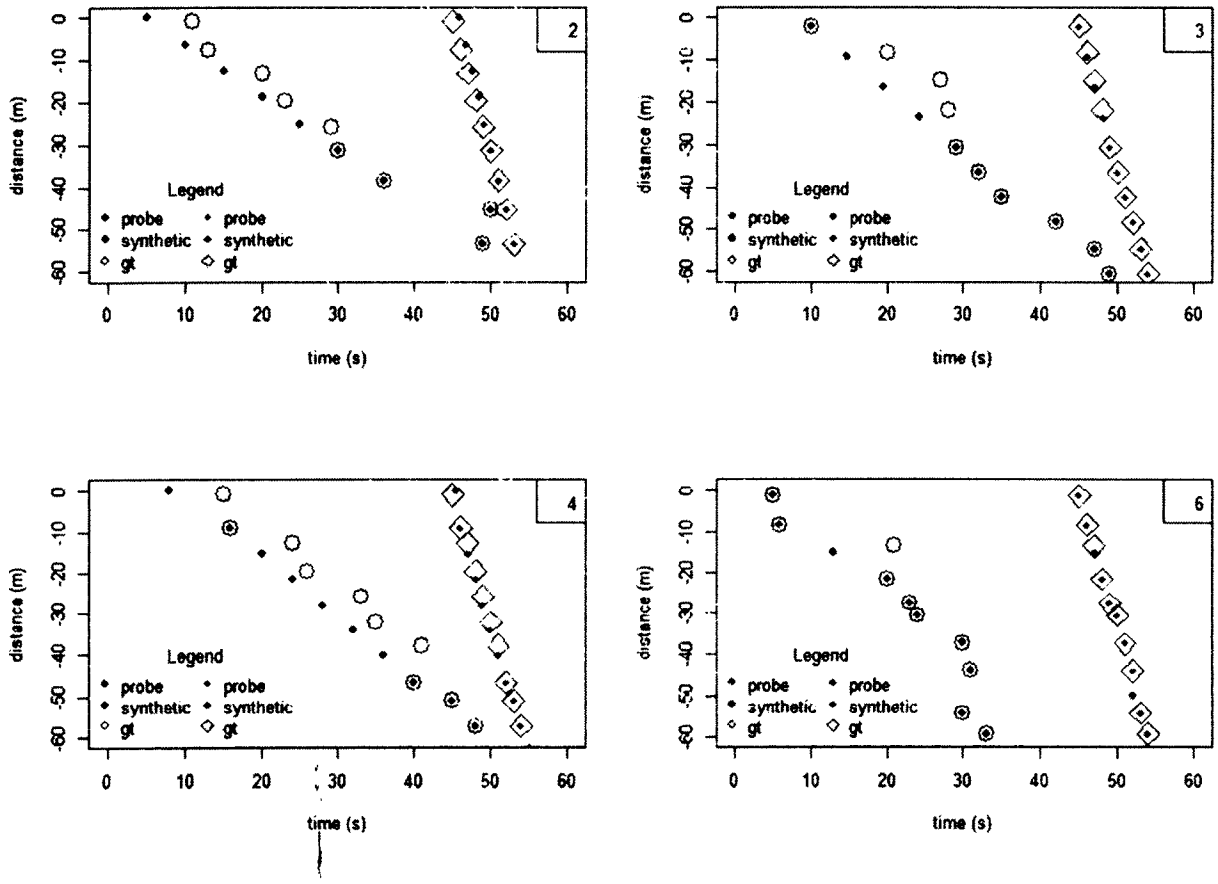


Figure 21. Synthetic, probe and ground truth vehicles' queuing and discharging points

The figure confirms that this assumption is relatively true for the discharging points, but it is not necessarily true for the queuing points. Therefore this become a source of error; however, this approach is the best that can be done since there is no other information available.

Spacing headway is an essential input to reconstruct the trajectories. How many vehicles can fit between two probe vehicles genuinely depends on spacing headway, particularly queuing spacing headway. The spacing headway calculated from vehicles stopping coordinates extracted

from training data was used to determine the mean spacing headway. The length of the car as documented in *Vissim* is 3.75 to 4.76 meters; therefore, the spacing data were subset for only the ones that were more than 3 meters and less than 10 meters, which is most likely to be the headway spacing for single vehicle.

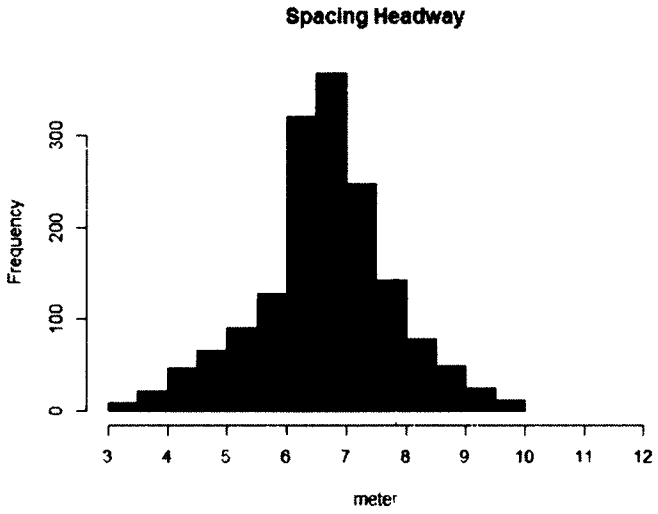


Figure 22. Spacing headway distribution

Figure 22 shows the spacing headway distribution, the mean space headway is 6.6 meters. In this study the mean space headway is determined to be 6 meters.

5.5.2 Acceleration and deceleration modeling

Table 7 shows that the best fit for the deceleration data for all scenarios is the logarithmic model, on the other hand the polynomial model is the best fit for the acceleration data for all models.

Table 7. The R^2 of linear and nonlinear model (acceleration and deceleration)

Model	R^2	
	Dec	Acc
Linear	0.29	0.24
Quadratic	0.29	0.24
Logarithmic	[redacted]	0.52
Exponential	0.16	0.17
Power	0.19	0.44
Polynomial	0.38	[redacted]

Based on this result, the logarithmic model was chosen to calculate the speed when the synthetic vehicle decelerates, likewise the polynomial model was selected for the synthetic vehicle acceleration.

Table 8. The acceleration and deceleration model

Acceleration type	Model
Deceleration (logarithmic model)	$speed = 48.36 - 10.62 \log(time)$
Acceleration (polynomial model)	$speed = -0.68 + 10.94\,(time) - 0.72\,(time)^2 + 0.01(time)^3$

Table 8 presents the acceleration and deceleration model built using the chosen regression model. The deceleration data indicated that the average duration for a vehicle to decelerate is 14 seconds while the acceleration data revealed that the average acceleration duration is 7 seconds. Using the model in Table 8, the synthetic vehicle speed when decelerate and accelerate were calculated. These speeds, combined with the cruising and stop speeds, were then used to form a synthetic speed profile as shown in Figure 23.

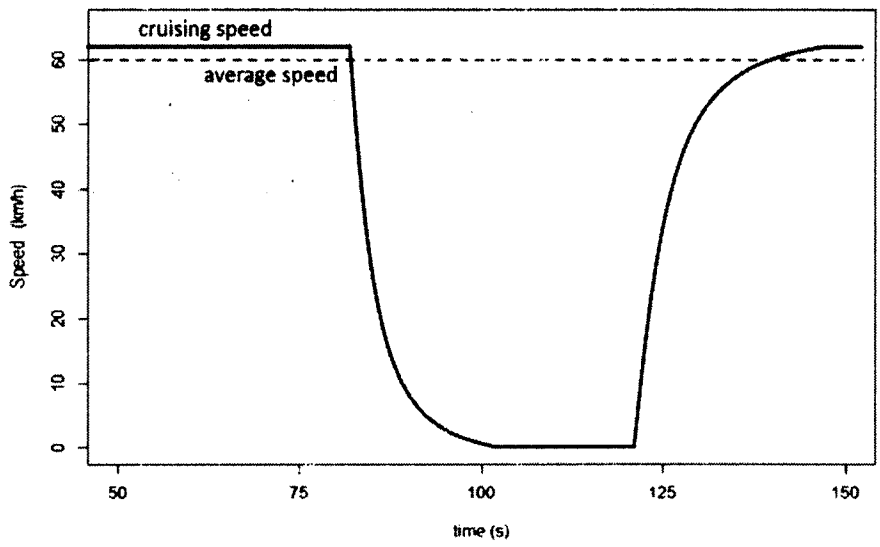


Figure 23. Illustration of the synthetic speed profile

Figure 23 gives an example of the synthetic speed profile. The cruising speed is the free flow speed, while in this study the average speed is taken from the desired speed in the simulation (60 km/h). This study is for undersaturated traffic condition cases and thus the vehicle is assumed to experience merely a single stop in the link.

5.5.3 The precision of fuel consumption estimation

Mean absolute error (MAE) was utilized to measure the bias between the fuel consumption estimation and the ground truth. The MAE are calculated according the formula in equation (49).

Similar to the previous research in this dissertation, the probe market penetration were varied from 10% to 100% by increment of 10%. The more the market penetration, the more the number of available cycles expected.

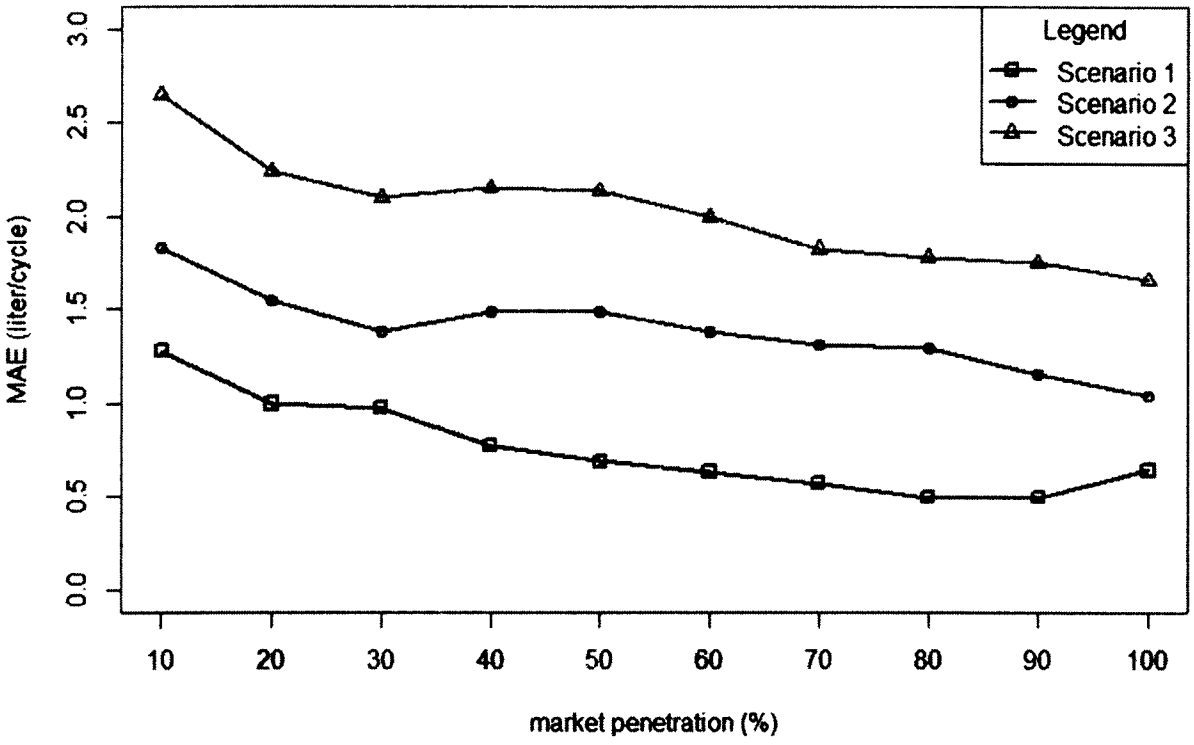


Figure 24. FC estimation MAE

Figure 24 shows that there is increase in error as the demand increases. On average the fuel consumption in scenario 2 generated about 0.63 liter more error compare to scenario 1; likewise the scenario 3 generated about 0.64 liter more error compare to scenario 2. The main reason for this is the increasing demand as also can be seen in the lane identification result. The accuracy of fuel consumption estimation more or less depends on the precision of the lane identification that has the same error tendency. Furthermore, the more vehicles (as in scenario 2 compared to scenario 1, and scenario 3 compared to scenario 2) the more variation can appear in each cycle and thus also becomes another source of error.

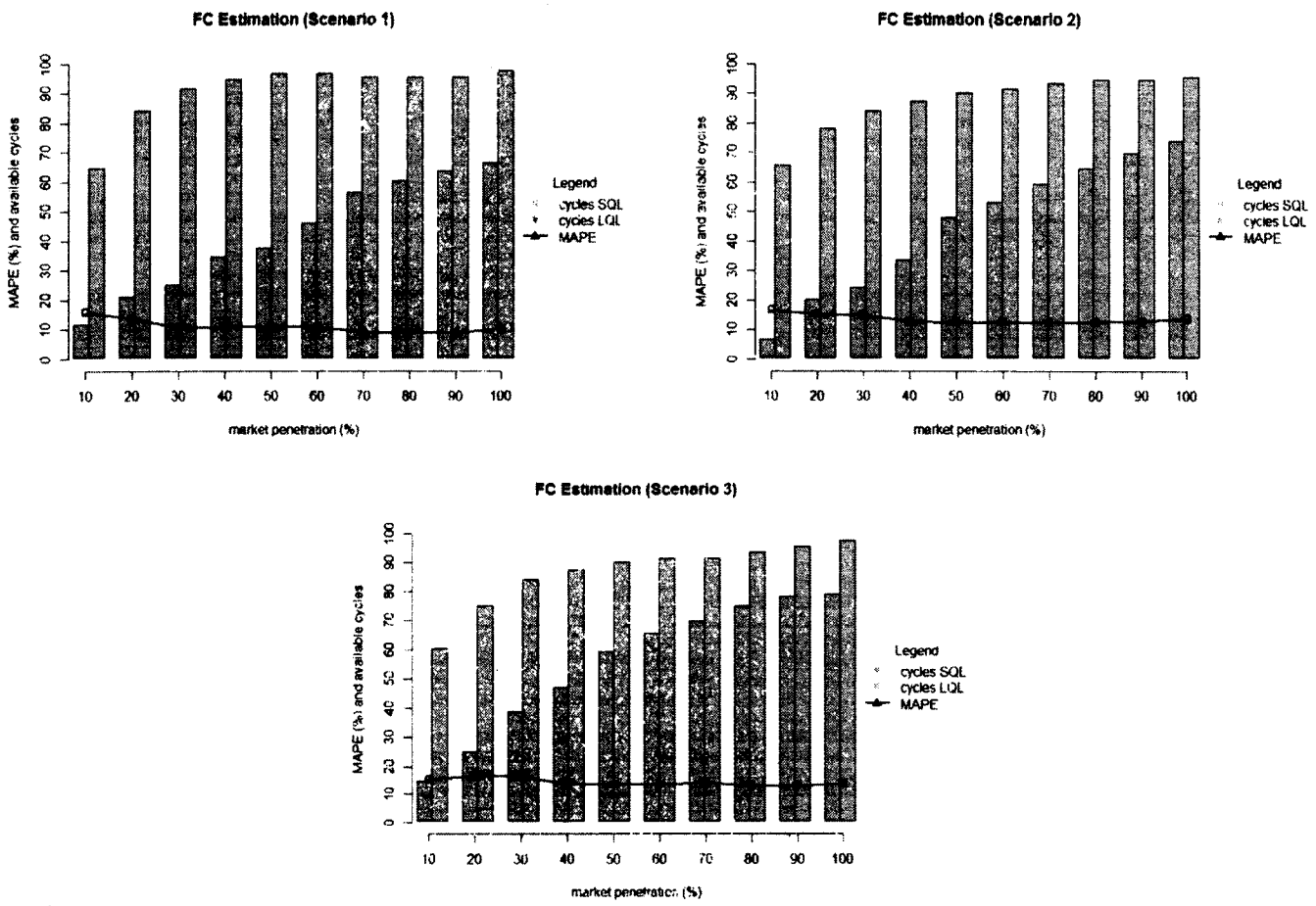


Figure 25. FC estimation MAPE and cycles available

The MAE values were compared to the fuel consumption mean value to understand the error relativity (mean absolute percentage error/MAPE) as shown in Figure 25. The figure shows that in scenario 1 the MAPE has reached less than or equal to 15% starting from 20% market penetration rate with 20 available cycles for SQL and 81 for LQL, likewise in scenarios 2 and 3 the MAE also are less than or equal to 15% starting from 40% market penetration where there are 32 available cycles for SQL, 84 for LQL and 45 for SQL, 84 for LQL, respectively. As expected, the more the market penetration, the less the MAE, and directly proportional to that, the MAPE. The full description of the fuel consumption MAE and available cycles can be seen

in Appendix T. The result presented here is the result without the missing cycle. As was discussed earlier, there will be missing cycles when using the probe vehicle especially in the low market penetration rate.

5.6 Summary and Conclusion

In this research, the real time fuel consumption estimation was conducted using probe vehicles by considering their lane position in the queue. The probe vehicle queuing and discharging information was used to build a shockwave profile for each cycle. Using this profile, synthetic trajectories were built. This synthetic trajectories became the base for assigning synthetic speed profile. There are four speed conditions for the profile which are cruising, decelerating, stopping and accelerating. The cruising speed was using the free flow speed while the stopping speed is equal to zero. The decelerating and accelerating speed were modeled using training data.

These synthetic profile speeds were then assigned to the reconstructed trajectories to be able to calculate the fuel consumption. Eventually, the fuel consumption of both probe and synthetic vehicles were calculated using the Virginia Tech Comprehensive Power-Based Fuel Consumption Model. The results are very encouraging, where the mean absolute error was satisfactory starting from a reasonable market penetration rate.

This research has demonstrated that the probe vehicles can be used properly to estimate the fuel consumption, particularly in signalized intersections where there are cases of unequal queues. The result shows that the more the arrival rates, the more variation appears which generates more error. Furthermore, this study has demonstrated that the precision of the lane identification process has a great impact on the fuel emissions accuracy. A limitation of this study is that the result displayed is only for the available cycles since there is no treatment for the missing cycle, which commonly arises when using probe vehicles. This issue can become a gap to be

considered in future study. Additionally, this study simplifies the vehicle types by assuming that they were all uniform. Future studies can be conducted by considering this issue to make this method more applicable.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusion

This study presents the use of probe vehicles for analyzing queue dynamics including fuel consumption estimation in signalized intersections. This is done by considering the probe vehicles' lane position particularly in the case of unequal queues at multiple-lane links. The conclusion can be divided in three categories: lane identification, queue length estimation and fuel consumption estimation.

6.1.1 Probe vehicles' lane identification

This study develops a framework for identifying the probe vehicles' lane in real time at signalized intersections in the case of unequal queues at multiple-lane signalized intersections. The proposed model is intended to be used as probe vehicle data preprocessing in estimating queue length and fuel consumption properly. The results demonstrate that:

1. Instead of using the queuing shockwave speed calculated from all probes vehicles available, the probe vehicles' individual shockwave speed can be employed to better identify their lane while queued.
2. Beside the individual queuing shockwave speed, the lane position of probe vehicles can also be characterized by their mean distance to the traffic light.
3. The probe vehicles' lane can be predicted properly when their individual queuing shockwave speed and distance predicted probability is larger or equal to 85% for short queue lane and larger or equal to 75% for long queue lane.

4. The more the arrival rates, the less the precision of the lane identification.

6.1.2 Queue length estimation

This study investigates the role of lane identification for estimating the queue length. Furthermore, this study explores the acceptable market penetration rate for queue length estimation using probe vehicles' information. The conclusions are:

1. The precision of probe vehicles' lane identification have a great impact for queue length estimation. The queue length accuracy is positively correlated to the lane identification precision.
2. The queue length estimation can be performed successfully by recognizing the last probe vehicle that its lane was identified properly.
3. The acceptable market penetration rate is 40% where in this rate the queue length estimation reached about 90% accuracy.

6.1.3 Fuel consumption estimation

This study presents methodology to utilize probe vehicles' information in fuel consumption estimation. After identifying the probe vehicles' lane, synthetic trajectories were reconstructed using shockwave profiles built from the probe vehicles' queuing information. The acceleration and deceleration modeling are carried out using training data. The result combined with the synthetic trajectories are used to produce a synthetic speed profile. Finally, the fuel consumption estimation using the probe vehicle and the synthetic trajectories are estimated using the Virginia Tech Comprehensive Power-Based Fuel Consumption Model. The results shows that:

1. The probe vehicles' information can be used to estimate the fuel consumption in signalized intersection properly.

2. The higher the arrival rate on SQL, the larger the estimation errors.
3. Similar to the queue length estimation, the lane identification process plays an important role in fuel consumption estimation.
4. The market penetration rate of 40% was adequate to produce acceptable results.

6.2 Recommendations for Further Research

In general, this research has demonstrated that the probe vehicles' lane identification plays an important role in queue length and fuel consumption estimation at signalized intersection. However, the following areas are recommended to improve the current research work:

1. Future research should consider the issue of heterogeneity in vehicles' type.
2. Future research needs to deal appropriately with the missing cycle as the result of the probe vehicles' low market penetration rate.

REFERENCES

1. Turner, S.M., et al., *Travel Time Data Collection Handbook.* 1998, Texas Transportation Institute Texas A&M University System: Office of Highway Information Management Federal Highway Administration U.S. Department of Transportation.
2. Chen, M. and S.I.J. Chien. *Dynamic Freeway Travel Time Prediction Using Probe Vehicle Data: Link-based vs. Path-based.* in *Transportation Research Board 80th Annual Meeting.* 2001. Washington, D.C.
3. Chien, S.I.-J. and C.M. Kuchipudi, *Dynamic Travel Time Prediction with Real-Time and Historic Data.* Journal of Transportation Engineering, 2003.
4. Chu, L., J.-S. Oh, and W. Recker. *Adaptive Kalman Filter Based Freeway Travel time Estimation.* in *2005 TRB Annual Meeting.* 2005.
5. Comert, G. and M. Cetin, *Queue length estimation from probe vehicle location and the impacts of sample size.* European Journal of Operational Research, 2009. 197(1): p. 196-202.
6. Dia, H. and K. Thomas, *Development and evaluation of arterial incident detection models using fusion of simulated probe vehicle and loop detector data.* Information Fusion, 2011. 12(1): p. 20-27.
7. Sethi, V., et al., *Arterial Incident Detection Using Fixed Detector and Probe Vehicle Data.* Transportation Research Part C, 1995. 3(2): p. 99-112.
8. Rakha, M.C.H. *Estimating Fuel Consumption and Carbon Footprint at Signalized Intersections using Probe Vehicle Trajectories.* in *TRB 2014 Annual Meeting.* 2014.
9. Chen, M. and S.I.J. Chien, *Determining the Number of Probe Vehicles for Freeway Travel Time Estimation by Microscopic Simulation.* Transportation Research Record 2000. 1719(1): p. 61-68.
10. Cetin, M., G.F. List, and Y. Zhou, *Factors Affecting Minimum Number of Probes Required for Reliable Estimation of Travel Time.* Transportation Research Record 2005. 1917(1): p. 37-44.
11. Liu, H.X. and W. Ma, *A virtual vehicle probe model for time-dependent travel time estimation on signalized arterials.* Transportation Research Part C: Emerging Technologies, 2009. 17(1): p. 11-26.
12. Xuan, Y. and B. Coifman. *Lane Change Maneuver Detection from Probe Vehicle DGPS Data.* in *IEEE Intelligent Transportation Systems Conference.* 2006. Toronto, Canada.
13. Nanthawichit, C., T. Nakatsuji, and H. Suzuki. *Application of Probe Vehicle Data for Real-Time Traffic State Estimation and Short-Term Travel Time Prediction on a Freeway.* in *TRB 2003 Annual Meeting.* 2003. Washington, D.C.
14. Dai, X., M.A. Ferman, and R.P. Roesser. *A Simulation Evaluation of a Real-Time Traffic Information System Using Probe Vehicles.* in *2003 IEEE Intelligent Transportation Systems.* 2003.
15. Cetin, M., *Estimating Queue Dynamics at Signalized Intersections from Probe Vehicle Data.* Transportation Research Record: Journal of the Transportation Research Board, 2012. 2315(-1): p. 164-172.
16. Comert, G., *Effect of stop line detection in queue length estimation at traffic signals from probe vehicles data.* European Journal of Operational Research, 2013. 226(1): p. 67-76.

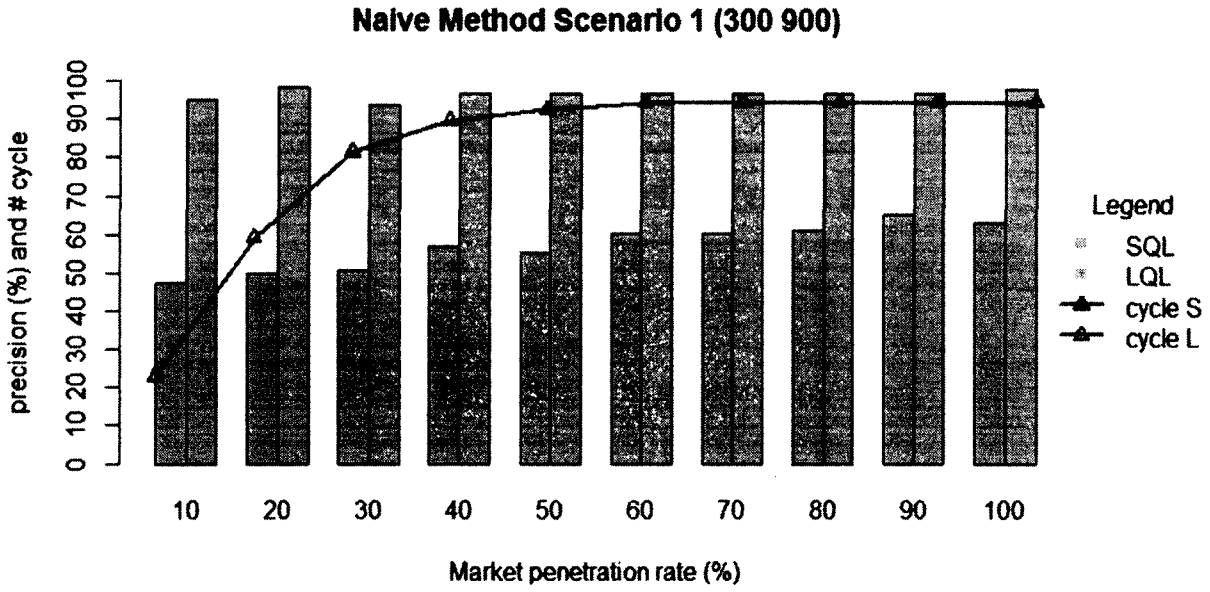
17. Ramezani, M. and N. Geroliminis, *Exploiting Probe Data to Estimate the Queue Profile in Urban Networks*, in *Proceedings of the 16th International IEEE Annual Conference on Intelligent Transportation Systems (ITSC 2013)*. 2013: The Hague, The Netherlands.
18. Antoniou, C., M. Ben-Akiva, and H.N. Koutsopoulos, *Incorporating Automated Vehicle Identification Data into Origin–Destination Estimation.* Transportation Research Record, 2004. 1882(1): p. 37–44.
19. Center, M.-A.U.T., *Investigation of the Implementation of a Probe-Vehicle Based Pavement Roughness Estimation System.* 2011.
20. Dawkins, B. and B. Powell, *Investigation of Pavement Maintenance Applications of Intellidrive SM (Final Report): Implementation and Deployment Factors for Vehicle Probe-based Pavement Maintenance (PBPM)*. 2011, May.
21. Brackstone, M., G. Fisher, and M. McDonald. *The use of probe vehicles on motorways, some empirical observations*. in *Proc. of 8th World Congress on ITS, Sydney, Australia.* 2001.
22. Srinivasan, K.K. and P.P. Jovanis, *Determination of number of probe vehicles required for reliable travel time measurement in urban network.* Transportation Research Record: Journal of the Transportation Research Board, 1996. 1537(1): p. 15-22.
23. Long Cheu, R., C. Xie, and D.H. Lee, *Probe vehicle population and sample size for arterial speed estimation.* Computer-Aided Civil and Infrastructure Engineering, 2002. 17(1): p. 53-60.
24. Zheng, F. and H. van Zuylen, *Estimating Urban Travel Time Distribution Using Probe Vehicle Data.* 2010.
25. Jenelius, E. and H.N. Koutsopoulos, *Travel time estimation for urban road networks using low frequency probe vehicle data.* Transportation Research Part B: Methodological, 2013. 53: p. 64-81.
26. Herring, R., et al. *Estimating arterial traffic conditions using sparse probe data*. in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on.* 2010. IEEE.
27. Qiu, T.Z., et al., *Estimation of freeway traffic density with loop detector and probe vehicle data.* Transportation Research Record: Journal of the Transportation Research Board, 2010. 2178(1): p. 21-29.
28. Kwon, J., K. Petty, and P. Varaiya, *Probe Vehicle Runs or Loop Detectors?: Effect of Detector Spacing and Sample Size on Accuracy of Freeway Congestion Monitoring.* Transportation Research Record: Journal of the Transportation Research Board, 2007. 2012(1): p. 57-63.
29. Hofleitner, A., et al., *Learning the dynamics of arterial traffic from probe data using a dynamic Bayesian network.* Intelligent Transportation Systems, IEEE Transactions on, 2012. 13(4): p. 1679-1693.
30. Shladover, S.E. and T.M. Kuhn, *Traffic probe data processing for full-scale deployment of vehicle-infrastructure integration.* Transportation Research Record: Journal of the Transportation Research Board, 2008. 2086(1): p. 115-123.
31. Ahn, K., et al , *Estimating vehicle fuel consumption and emissions based on instantaneous speed and acceleration levels.* Journal of Transportation Engineering, 2002. 128(2): p. 182-190.

32. Vieira da Rocha, T., et al., *Are vehicle trajectories simulated by dynamic traffic models relevant for estimating fuel consumption?* Transportation Research Part D: Transport and Environment, 2013. 24: p. 17-26.
33. Lighthill, M.J. and G.B. Whitham, *On Kinematic Waves. II. A Theory of Traffic Flow on Long Crowded Roads.* Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 1955. 229(1178): p. 317-345.
34. Richards, P.I., *Shock Waves on the Highway.* Operations Research, 1956. 4(1): p. 42-51.
35. Akcelik, R., *An interpretation of the parameters in the simple average travel speed model of fuel consumption.* Australian Road Research, 1985. 15(1).
36. Yue, H., *Mesoscopic fuel consumption and emission modeling.* 2008.
37. Evans, L. and R. Herman, *Automobile fuel economy on fixed urban driving schedules.* Transportation Science, 1978. 12(2): p. 137-152.
38. Rakha, H., K. Ahn, and A. Trani, *Comparison of MOBILE5a, MOBILE6, VT-MICRO, and CMEM models for estimating hot-stabilized light-duty gasoline vehicle emissions.* Canadian Journal of Civil Engineering, 2003. 30(6): p. 1010-1021.
39. Scora, G. and M. Barth, *Comprehensive Modal Emission Model (CMEM), Version 3.01 l User's Guide. University of California, Riverside.* Center for Environmental Research and, 2006. 2.
40. Chang, M.-F. and R. Herman, *Trip time versus stop time and fuel consumption characteristics in cities.* Transportation Science, 1981. 15(3): p. 183-209.
41. Watson, H., E. Milkins, and G. Marshall. *A simplified method for quantifying fuel consumption of vehicles in urban traffic.* in *Automotive Engineering Conference, 4th, 1979, Melbourne, Australia.* 1979.
42. Ahn, K., *Microscopic fuel consumption and emission modeling.* 1998, Virginia Polytechnic Institute and State University.
43. Rakha, H., et al., *Requirements for evaluating traffic signal control impacts on energy and emissions based on instantaneous speed and acceleration measurements.* Transportation Research Record: Journal of the Transportation Research Board, 2000. 1738(1): p. 56-67.
44. *Motor Vehicle Emission Simulator (MOVES), User Guide for MOVES2010b* U.S.E.P.A. Office of Transportation and Air Quality, Editor. 2012, United States Environmental Protection Agency.
45. Vallamsundar, S. and J. Lin, *MOVES Versus MOBILE.* Transportation Research Record: Journal of the Transportation Research Board, 2011. 2233(-1): p. 27-35.
46. Akçelik, R. *Progress in fuel consumption modelling for urban traffic management.* 1983. Australian Road Research Board.
47. Bachman, W., W. Sarasua, and R. Guensler, *Geographic information system framework for modeling mobile-source emissions.* Transportation Research Record: Journal of the Transportation Research Board, 1996. 1551(1): p. 123-132.
48. Bachman, W., et al., *Modeling regional mobile source emissions in a geographic information system framework.* Transportation Research Part C: Emerging Technologies, 2000. 8(1): p. 205-229.
49. Hunter, M.P., et al., *A Probe-Vehicle-Based Evaluation of Adaptive Traffic Signal Control.* IEEE Transactions on Intelligent Transportation Systems, 2012. 13(2).

50. Cheng, Y., et al., *Cycle-by-Cycle Queue Length Estimation for Signalized Intersections Using Sampled Trajectory Data.* Transportation Research Record: Journal of the Transportation Research Board, 2011. 2257(-1): p. 87-94.
51. Hao, P., X. Ban, and J. Whon Yu, *Kinematic Equation-Based Vehicle Queue Location Estimation Method for Signalized Intersections Using Mobile Sensor Data.* Journal of Intelligent Transportation Systems, 2014: p. 1-17.
52. Badillo, B.E., et al., *Queue Length Estimation using Conventional Vehicle Detector and Probe Vehicle Data*, in *2012 15th International IEEE Conference on Intelligent Transportation Systems* 2012: Anchorage, Alaska, USA.
53. Ban, X., P. Hao, and Z. Sun, *Real time queue length estimation for signalized intersections using travel times from mobile sensors.* Transportation Research Part C: Emerging Technologies, 2011. 19(6): p. 1133-1156.
54. Comert, G. and M. Cetin, *Analytical Evaluation of the Error in Queue Length Estimation at Traffic Signals From Probe Vehicle Data.* IEEE Transactions on Intelligent Transportation Systems, 2011. 12(2).
55. PTV, *Vissim 6.0 User Manual.* 2014.
56. Liu, H.X., et al., *Real-time queue length estimation for congested signalized intersections.* Transportation Research Part C: Emerging Technologies, 2009. 17(4): p. 412-427.
57. Vigos, G., M. Papageorgiou, and Y. Wang, *Real-time estimation of vehicle-count within signalized links.* Transportation Research Part C: Emerging Technologies, 2008. 16(1): p. 18-35.
58. Wu, J., X. Jin, and A.J. Horowitz, *Methodologies for Estimating Vehicle Queue Length at Metered On-Ramps.* Transportation Research Record: Journal of the Transportation Research Board, 2008. 2047(-1): p. 75-82.
59. Chang, G.-L. and C.-C. Su, *Predicting Intersection Queue With Neural Network Models.* Transportation Research Part C 1995. 3(3): p. 175-191.
60. Geroliminis, N. and A. Skabardonis, *Prediction of Arrival Profiles and Queue Lengths Along Signalized Arterials by Using a Markov Decision Process.* Transportation Research Record: Journal of the Transportation Research Board, 2005(1934): p. 116–124.
61. Viti, F. and H.J.v. Zuylen, *Modeling Queues at Signalized Intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2004(1883): p. 68–77.
62. Oosterbaan, R.J. and H. Ritzema, *Frequency and regression analysis.* Drainage principles and applications., 1994(Ed. 2): p. 175-223.
63. Fraley, C. and A.E. Raftery, *Model-based methods of classification: Using the mclust software in chemometrics.* Journal of Statistical Software, 2007. 18(6): p. 1-13.
64. Fraley, C., et al., *Package 'mclust'.* 2015.
65. Herrera, J.C., et al., *Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment.* Transportation Research Part C: Emerging Technologies, 2010. 18(4): p. 568-583.
66. Hoh, B., et al., *Enhancing security and privacy in traffic-monitoring systems.* Pervasive Computing, IEEE, 2006. 5(4): p. 38-46.
67. Krumm, J., *Inference attacks on location tracks*, in *Pervasive Computing*. 2007, Springer. p. 127-143.

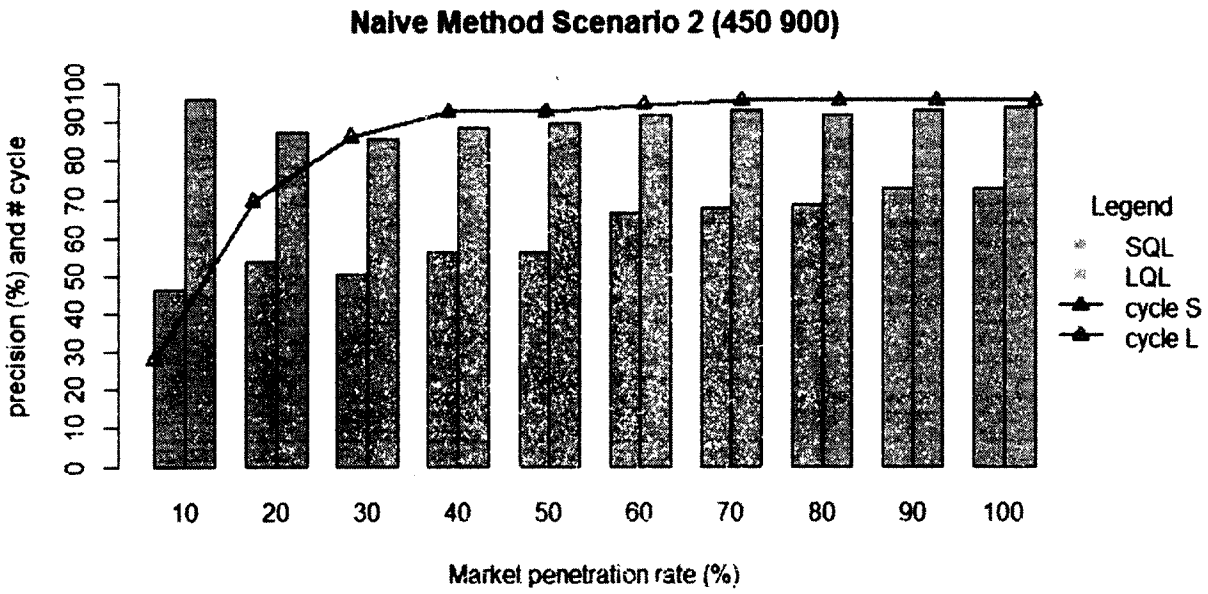
68. Krause, A., et al. *Toward community sensing.* in *Proceedings of the 7th international conference on Information processing in sensor networks.* 2008. IEEE Computer Society.
69. Lee, S., S. Wong, and Y. Li, *Real-time estimation of lane-based queue lengths at isolated signalized junctions.* Transportation Research Part C: Emerging Technologies, 2015. 56: p. 1-17.
70. Sharma, A., D. Bullock, and J. Bonneson, *Input-output and hybrid techniques for real-time prediction of delay and maximum queue length at signalized intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2007(2035): p. 69-80.
71. Lawson, T., D. Lovell, and C. Daganzo, *Using input-output diagram to determine spatial and temporal extents of a queue upstream of a bottleneck.* Transportation Research Record: Journal of the Transportation Research Board, 1997(1572): p. 140-147.
72. Wu, X. and H.X. Liu, *A shockwave profile model for traffic flow on congested urban arterials.* Transportation Research Part B: Methodological, 2011. 45(10): p. 1768-1786.
73. Caiazzo, F., et al., *Air pollution and early deaths in the United States. Part I: Quantifying the impact of major sectors in 2005.* Atmospheric Environment, 2013. 79: p. 198-208.
74. EPA, *Mobile 5A user guide*, U.S.E.P. Agency, Editor. 1993: Ann Arbor, Mich.
75. EPA, *Draft user's guide to MOBILE6.0, Mobile source emission factor model*, U.S.E.P. Agency, Editor. 2001: Ann Arbor, Mich.
76. Cappiello, A., et al. *A statistical model of vehicle emissions and fuel consumption.* in *Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference on.* 2002. IEEE.
77. Smit, R., M. Poelman, and J. Schrijver, *Improved road traffic emission inventories by adding mean speed distributions.* Atmospheric Environment, 2008. 42(5): p. 916-926.
78. Andre, M. and U. Hammarstrom, *Driving speeds in Europe for pollutant emissions estimation.* Transportation Research Part D, 2000. 5: p. 321-335.
79. Rakha, H.A., et al., *Virginia tech comprehensive power-based fuel consumption model: model development and testing.* Transportation Research Part D: Transport and Environment, 2011. 16(7): p. 492-503.

APPENDIX A - LANE IDENTIFICATION USING THE NAÏVE METHOD (SCENARIO 1)



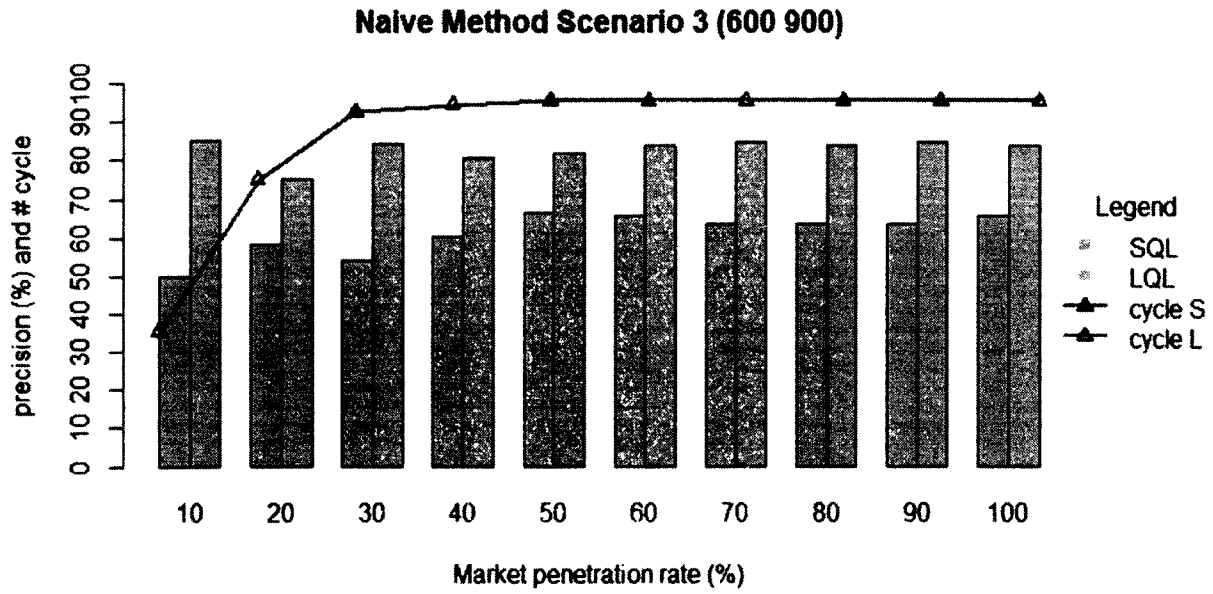
p	Precision		Cycles		Precision (Last Probe)		Cycles (Last Probe)	
	SQL	LQL	SQL	LQL	SQL	LQL	SQL	LQL
0.1	48%	94%	21	21	48%	95%	21	21
0.2	50%	97%	60	60	50%	98%	60	60
0.3	49%	95%	84	84	51%	94%	84	84
0.4	51%	96%	93	93	57%	97%	93	93
0.5	53%	95%	96	96	55%	97%	96	96
0.6	54%	95%	98	98	60%	97%	98	98
0.7	58%	95%	98	98	60%	97%	98	98
0.8	59%	95%	98	98	61%	97%	98	98
0.9	59%	95%	98	98	65%	97%	98	98
1	57%	95%	98	98	63%	98%	98	98

APPENDIX B - LANE IDENTIFICATION USING THE NAÏVE METHOD (SCENARIO 2)



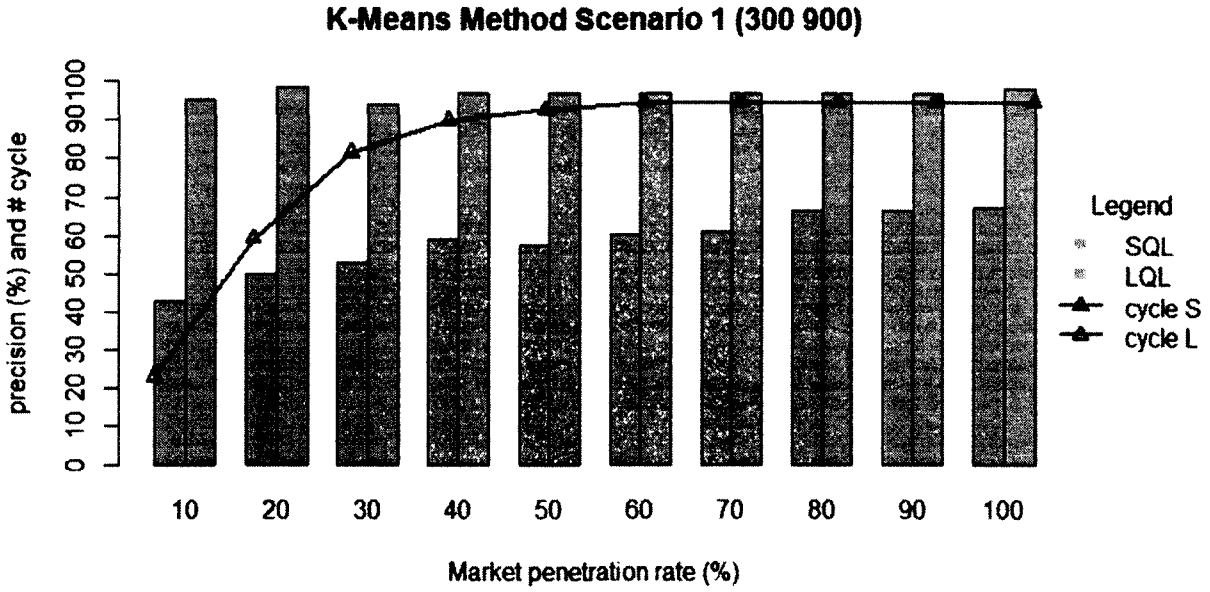
p	Precision		Cycles		Precision (Last Probe)		Cycles (Last Probe)	
	SQL	LQL	SQL	LQL	SQL	LQL	SQL	LQL
0.1	41%	96%	26	26	46%	96%	26	26
0.2	50%	87%	71	71	54%	87%	71	71
0.3	49%	84%	89	89	51%	85%	89	89
0.4	54%	86%	96	96	56%	89%	96	96
0.5	55%	86%	96	96	56%	90%	96	96
0.6	57%	87%	98	98	66%	92%	98	98
0.7	58%	86%	99	99	68%	93%	99	99
0.8	57%	87%	99	99	69%	92%	99	99
0.9	58%	87%	99	99	73%	93%	99	99
1	60%	87%	99	99	73%	94%	99	99

APPENDIX C - LANE IDENTIFICATION USING THE NAÏVE METHOD (SCENARIO 3)



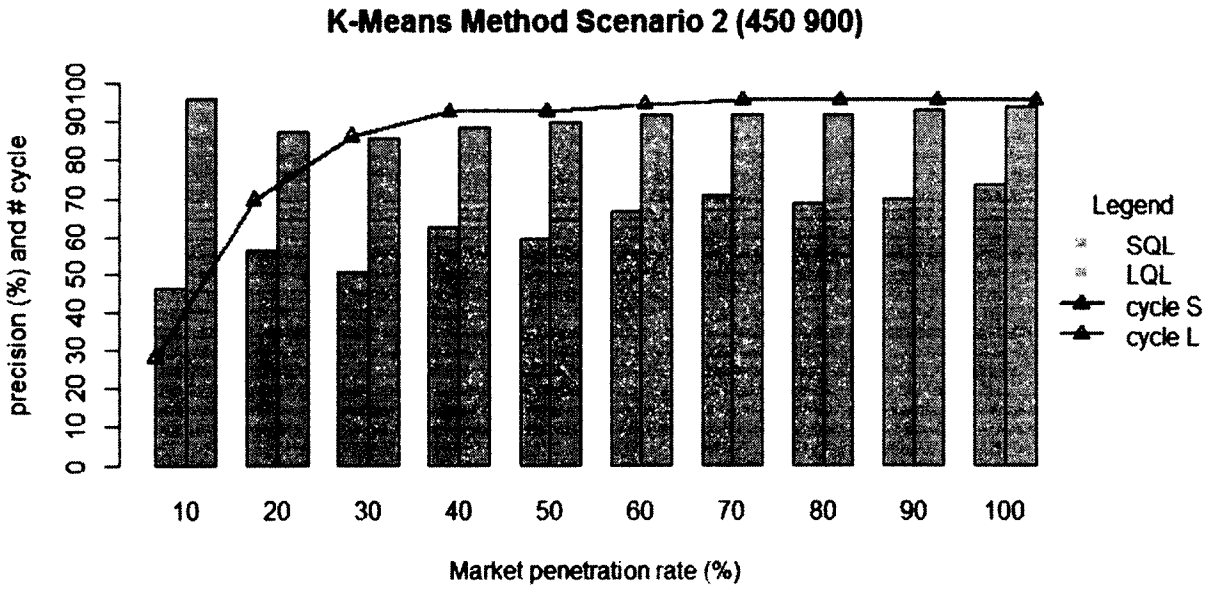
p	Precision		Cycles		Precision (Last Probe)		Cycles (Last Probe)	
	SQL	LQL	SQL	LQL	SQL	LQL	SQL	LQL
0.1	45%	81%	34	34	50%	85%	34	34
0.2	55%	74%	77	77	58%	75%	77	77
0.3	53%	74%	96	96	54%	84%	96	96
0.4	55%	74%	98	98	60%	81%	98	98
0.5	58%	74%	99	99	67%	82%	99	99
0.6	58%	74%	99	99	66%	84%	99	99
0.7	56%	74%	99	99	64%	85%	99	99
0.8	56%	74%	99	99	64%	84%	99	99
0.9	55%	74%	99	99	64%	85%	99	99
1	55%	73%	99	99	66%	84%	99	99

APPENDIX D - LANE IDENTIFICATION USING THE K-MEANS METHOD (SCENARIO 1)



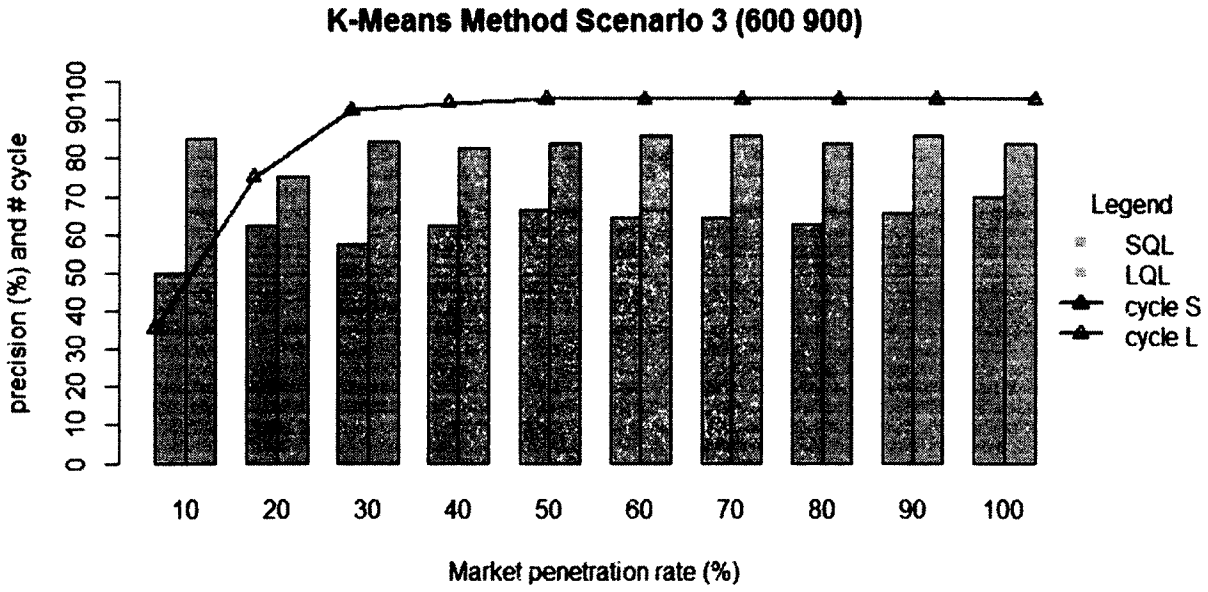
p	Precision		Cycles		Precision (Last Probe)		Cycles (Last Probe)	
	SQL	LQL	SQL	LQL	SQL	LQL	SQL	LQL
0.1	44%	91%	21	21	43%	95%	21	21
0.2	49%	96%	60	60	50%	98%	60	60
0.3	47%	95%	84	84	53%	94%	84	84
0.4	51%	96%	93	93	59%	97%	93	93
0.5	53%	95%	96	96	57%	97%	96	96
0.6	54%	95%	98	98	60%	97%	98	98
0.7	58%	95%	98	98	61%	97%	98	98
0.8	61%	95%	98	98	66%	97%	98	98
0.9	59%	95%	98	98	66%	97%	98	98
1	58%	96%	98	98	67%	98%	98	98

APPENDIX E - LANE IDENTIFICATION USING THE K-MEANS METHOD (SCENARIO 2)



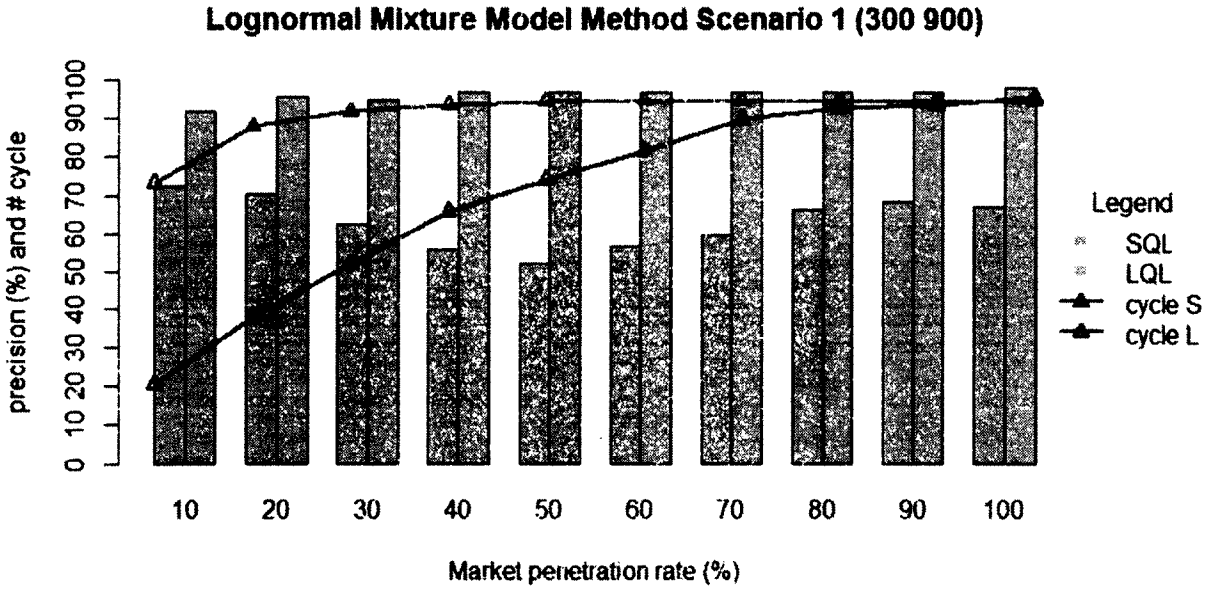
p	Precision		Cycles		Precision (Last Probe)		Cycles (Last Probe)	
	SQL	LQL	SQL	LQL	SQL	LQL	SQL	LQL
0.1	40%	95%	26	26	46%	96%	26	26
0.2	53%	87%	71	71	56%	87%	71	71
0.3	51%	86%	89	89	51%	85%	89	89
0.4	55%	88%	96	96	63%	89%	96	96
0.5	54%	88%	96	96	59%	90%	96	96
0.6	59%	88%	98	98	66%	92%	98	98
0.7	57%	88%	99	99	71%	92%	99	99
0.8	58%	88%	99	99	69%	92%	99	99
0.9	58%	88%	99	99	70%	93%	99	99
1	60%	89%	99	99	74%	94%	99	99

APPENDIX F - LANE IDENTIFICATION USING THE K-MEANS METHOD (SCENARIO 3)



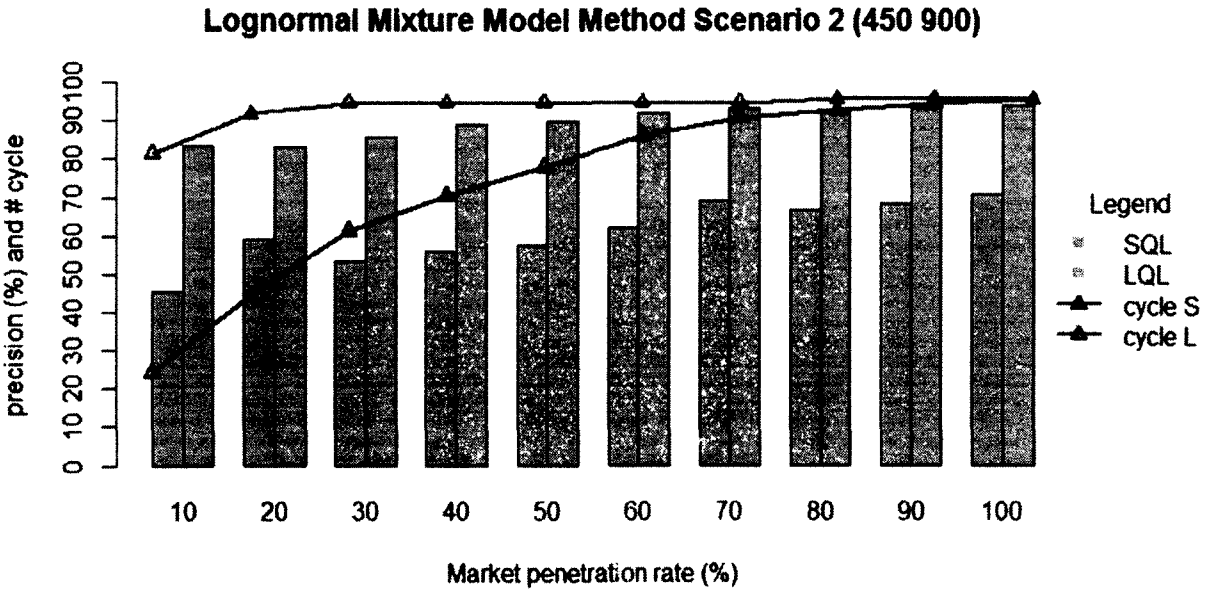
p	Precision		Cycles		Precision (Last Probe)		Cycles (Last Probe)	
	SQL	LQL	SQL	LQL	SQL	LQL	SQL	LQL
0.1	44%	80%	34	34	50%	85%	34	34
0.2	56%	74%	77	77	62%	75%	77	77
0.3	55%	75%	96	96	57%	84%	96	96
0.4	55%	76%	98	98	62%	83%	98	98
0.5	57%	76%	99	99	67%	84%	99	99
0.6	57%	76%	99	99	65%	86%	99	99
0.7	56%	75%	99	99	65%	86%	99	99
0.8	55%	75%	99	99	63%	84%	99	99
0.9	56%	75%	99	99	66%	86%	99	99
1	57%	75%	99	99	70%	84%	99	99

APPENDIX G - LANE IDENTIFICATION USING THE LOGNORMAL MODEL METHOD (SCENARIO 1)



p	Precision		Cycles		Precision (Last Probe)		Cycles (Last Probe)	
	SQL	LQL	SQL	LQL	SQL	LQL	SQL	LQL
0.1	68%	86%	18	75	72%	92%	18	75
0.2	67%	85%	37	91	70%	96%	37	91
0.3	61%	85%	52	95	62%	95%	52	95
0.4	55%	85%	67	97	56%	97%	67	97
0.5	52%	84%	76	98	52%	97%	76	98
0.6	53%	83%	84	98	56%	97%	84	98
0.7	58%	83%	93	98	60%	97%	93	98
0.8	58%	82%	96	98	66%	97%	96	98
0.9	58%	83%	97	98	68%	97%	97	98
1	58%	83%	99	98	67%	98%	99	98

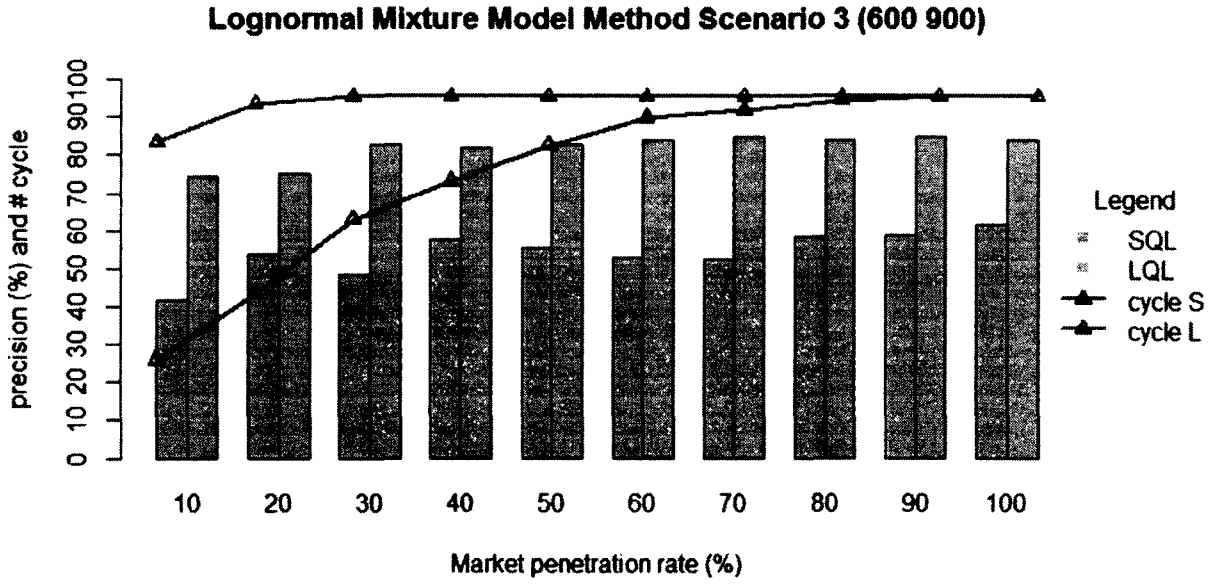
APPENDIX H - LANE IDENTIFICATION USING THE LOGNORMAL MODEL METHOD (SCENARIO 2)



p	Precision		Cycles		Precision (Last Probe)		Cycles (Last Probe)	
	SQL	LQL	SQL	LQL	SQL	LQL	SQL	LQL
0.1	48%	80%	22	84	45%	83%	22	84
0.2	60%	77%	44	95	59%	83%	44	95
0.3	54%	76%	62	98	53%	86%	62	98
0.4	58%	76%	72	98	56%	89%	72	98
0.5	57%	76%	80	98	58%	90%	80	98
0.6	58%	77%	89	98	62%	92%	89	98
0.7	58%	76%	94	98	69%	93%	94	98
0.8	56%	76%	96	99	67%	92%	96	99
0.9	56%	76%	98	99	68%	93%	98	99
1	57%	76%	99	99	71%	94%	99	99

APPENDIX I - LANE IDENTIFICATION USING THE LOGNORMAL MODEL METHOD (SCENARIO 3)

p	Precision		Cycles		Precision (Last Probe)		Cycles (Last Probe)	
	SQL	LQL	SQL	LQL	SQL	LQL	SQL	LQL
0.1	46%	69%	24	86	42%	74%	24	86
0.2	56%	66%	43	97	53%	75%	43	97
0.3	51%	66%	64	99	48%	83%	64	99
0.4	56%	67%	75	99	57%	82%	75	99
0.5	56%	65%	85	99	55%	83%	85	99
0.6	55%	65%	93	99	53%	84%	93	99
0.7	54%	66%	95	99	53%	85%	95	99
0.8	56%	66%	98	99	58%	84%	98	99
0.9	56%	66%	99	99	59%	85%	99	99
1	56%	66%	99	99	62%	84%	99	99



APPENDIX J - LANE IDENTIFICATION RESULTS COMPARISON FOR SCENARIO 1

Method	p	Precision		Cycles		Precision (Last Probe)		Cycles (Last Probe)	
		SQL	LQL	SQL	LQL	SQL	LQL	SQL	LQL
Naïve	0.1	48%	94%	21	21	48%	95%	21	21
	0.2	50%	97%	60	60	50%	98%	60	60
	0.3	49%	95%	84	84	51%	94%	84	84
	0.4	51%	96%	93	93	57%	97%	93	93
	0.5	53%	95%	96	96	55%	97%	96	96
	0.6	54%	95%	98	98	60%	97%	98	98
	0.7	58%	95%	98	98	60%	97%	98	98
	0.8	59%	95%	98	98	61%	97%	98	98
	0.9	59%	95%	98	98	65%	97%	98	98
	1	57%	95%	98	98	63%	98%	98	98
K-Means	0.1	44%	91%	21	21	48%	95%	21	21
	0.2	49%	96%	60	60	50%	98%	60	60
	0.3	47%	95%	84	84	51%	94%	84	84
	0.4	51%	96%	93	93	57%	97%	93	93
	0.5	53%	95%	96	96	55%	97%	96	96
	0.6	54%	95%	98	98	60%	97%	98	98
	0.7	58%	95%	98	98	60%	97%	98	98
	0.8	61%	95%	98	98	61%	97%	98	98
	0.9	59%	95%	98	98	65%	97%	98	98
	1	58%	96%	98	98	63%	98%	98	98
Lognormal	0.1	68%	86%	18	75	72%	92%	18	75
	0.2	67%	85%	37	91	70%	96%	37	91
	0.3	61%	85%	52	95	62%	95%	52	95
	0.4	55%	85%	67	97	56%	97%	67	97
	0.5	52%	84%	76	98	52%	97%	76	98
	0.6	53%	83%	84	98	56%	97%	84	98
	0.7	58%	83%	93	98	60%	97%	93	98
	0.8	58%	82%	96	98	66%	97%	96	98
	0.9	58%	83%	97	98	68%	97%	97	98
	1	58%	83%	99	98	67%	98%	99	98
Bivariate	0.1	92%	95%	11	62	91%	98%	11	62
Normal	0.2	88%	93%	20	81	85%	98%	20	81
	0.3	90%	94%	24	88	92%	97%	24	88
	0.4	87%	94%	33	91	91%	98%	33	91
	0.5	83%	94%	36	93	86%	96%	36	93
	0.6	77%	94%	44	93	82%	97%	44	93
	0.7	81%	94%	54	92	85%	97%	54	92
	0.8	83%	94%	58	92	86%	97%	58	92
	0.9	83%	94%	61	92	87%	97%	61	92
	1	83%	94%	64	94	88%	98%	64	94

APPENDIX K - LANE IDENTIFICATION RESULTS COMPARISON FOR SCENARIO 2

Method	p	Precision		Cycles		Precision (Last Probe)		Cycles (Last Probe)	
		SQL	LQL	SQL	LQL	SQL	LQL	SQL	LQL
Naïve	0.1	41%	96%	26	26	46%	96%	26	26
	0.2	50%	87%	71	71	54%	87%	71	71
	0.3	49%	84%	89	89	51%	85%	89	89
	0.4	54%	86%	96	96	56%	89%	96	96
	0.5	55%	86%	96	96	56%	90%	96	96
	0.6	57%	87%	98	98	66%	92%	98	98
	0.7	58%	86%	99	99	68%	93%	99	99
	0.8	57%	87%	99	99	69%	92%	99	99
	0.9	58%	87%	99	99	73%	93%	99	99
	1	60%	87%	99	99	73%	94%	99	99
K-Means	0.1	40%	95%	26	26	46%	96%	26	26
	0.2	53%	87%	71	71	56%	87%	71	71
	0.3	51%	86%	89	89	51%	85%	89	89
	0.4	55%	88%	96	96	63%	89%	96	96
	0.5	54%	88%	96	96	59%	90%	96	96
	0.6	59%	88%	98	98	66%	92%	98	98
	0.7	57%	88%	99	99	71%	92%	99	99
	0.8	58%	88%	99	99	69%	92%	99	99
	0.9	58%	88%	99	99	70%	93%	99	99
	1	60%	89%	99	99	74%	94%	99	99
Lognormal	0.1	48%	80%	22	84	45%	83%	22	84
	0.2	60%	77%	44	95	59%	83%	44	95
	0.3	54%	76%	62	98	53%	86%	62	98
	0.4	58%	76%	72	98	56%	89%	72	98
	0.5	57%	76%	80	98	58%	90%	80	98
	0.6	58%	77%	89	98	62%	92%	89	98
	0.7	58%	76%	94	98	69%	93%	94	98
	0.8	56%	76%	96	99	67%	92%	96	99
	0.9	56%	76%	98	99	68%	93%	98	99
	1	57%	76%	99	99	71%	94%	99	99
Bivariate Normal	0.1	86%	92%	6	63	83%	94%	6	63
	0.2	86%	91%	19	75	84%	93%	19	75
	0.3	80%	91%	23	81	83%	93%	23	81
	0.4	80%	90%	32	84	88%	93%	32	84
	0.5	81%	89%	46	87	83%	95%	46	87
	0.6	78%	88%	51	88	80%	95%	51	88
	0.7	78%	88%	57	90	81%	94%	57	90
	0.8	78%	88%	62	91	81%	95%	62	91
	0.9	77%	89%	67	91	78%	95%	67	91
	1	77%	89%	71	92	79%	95%	71	92

APPENDIX L - LANE IDENTIFICATION RESULTS COMPARISON FOR SCENARIO 3

Method	p	Precision		Cycles		Precision (Last Probe)		Cycles (Last Probe)	
		SQL	LQL	SQL	LQL	SQL	LQL	SQL	LQL
Naïve	0.1	45%	81%	34	34	50%	85%	34	34
	0.2	55%	74%	77	77	58%	75%	77	77
	0.3	53%	74%	96	96	54%	84%	96	96
	0.4	55%	74%	98	98	60%	81%	98	98
	0.5	58%	74%	99	99	67%	82%	99	99
	0.6	58%	74%	99	99	66%	84%	99	99
	0.7	56%	74%	99	99	64%	85%	99	99
	0.8	56%	74%	99	99	64%	84%	99	99
	0.9	55%	74%	99	99	64%	85%	99	99
	1	55%	73%	99	99	66%	84%	99	99
K-Means	0.1	44%	80%	34	34	50%	85%	34	34
	0.2	56%	74%	77	77	62%	75%	77	77
	0.3	55%	75%	96	96	57%	84%	96	96
	0.4	55%	76%	98	98	62%	83%	98	98
	0.5	57%	76%	99	99	67%	84%	99	99
	0.6	57%	76%	99	99	65%	86%	99	99
	0.7	56%	75%	99	99	65%	86%	99	99
	0.8	55%	75%	99	99	63%	84%	99	99
	0.9	56%	75%	99	99	66%	86%	99	99
	1	57%	75%	99	99	70%	84%	99	99
Lognormal	0 1	46%	69%	24	86	42%	74%	24	86
	0.2	56%	66%	43	97	53%	75%	43	97
	0.3	51%	66%	64	99	48%	83%	64	99
	0.4	56%	67%	75	99	57%	82%	75	99
	0.5	56%	65%	85	99	55%	83%	85	99
	0.6	55%	65%	93	99	53%	84%	93	99
	0.7	54%	66%	95	99	53%	85%	95	99
	0.8	56%	66%	98	99	58%	84%	98	99
	0.9	56%	66%	99	99	59%	85%	99	99
	1	56%	66%	99	99	62%	84%	99	99
Bivariate	0.1	79%	84%	14	58	79%	88%	14	58
Normal	0.2	77%	83%	24	72	75%	88%	24	72
	0.3	71%	81%	37	81	73%	85%	37	81
	0.4	72%	81%	45	84	71%	83%	45	84
	0.5	75%	80%	57	87	77%	83%	57	87
	0.6	74%	81%	63	88	75%	83%	63	88
	0.7	72%	80%	67	88	75%	84%	67	88
	0.8	70%	80%	72	90	78%	83%	72	90
	0.9	72%	81%	75	92	79%	85%	75	92
	1	71%	80%	76	94	78%	83%	76	94

APPENDIX M - BIVARIATE MIXTURE MODEL DENSITY ESTIMATION

Lane 1 (SQL)

```
-------------------------------------------------------
Density estimation via Gaussian finite mixture modeling
-------------------------------------------------------

Mclust VVE (ellipsoidal, equal orientation) model with 5 components:

 log.likelihood   n df      BIC       ICL
      -898.5791 361 25 -1944.38 -1978.328

Clustering table:
  1   2   3   4   5
 90  60  52  30 129
```

Lane 2 (LQL)

```
-------------------------------------------------------
Density estimation via Gaussian finite mixture modeling
-------------------------------------------------------

Mclust VVV (ellipsoidal, varying volume, shape, and orientation) model with 8
components:

 log.likelihood    n df       BIC       ICL
       -6362.11 1348 47 -13062.92 -13756.06

Clustering table:
  1   2   3   4   5   6   7   8
 97  85  94 362 292 261  77  80
```

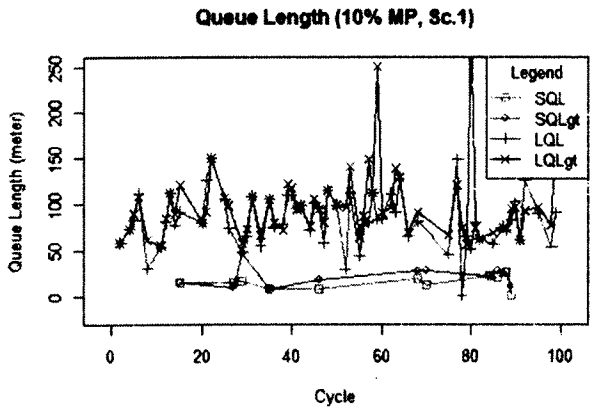
APPENDIX N - QUEUE LENGTH ESTIMATION BY APPLYING THE THRESHOLD FIRST

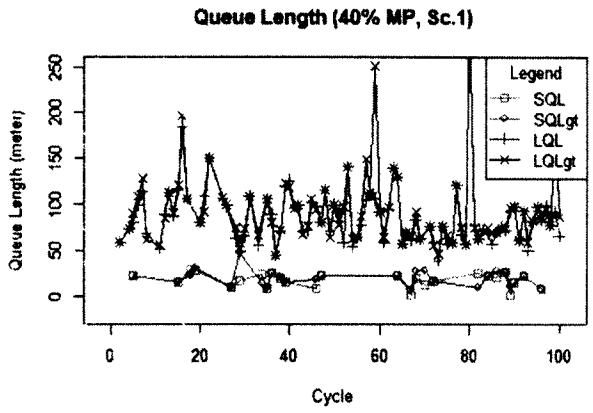
MP	MAE1	PerMAE1	MAE2	PerMAE2	ncycle1	ncycle2
Scenario 1						
0.1	7.7	35%	20.2	22%	11	62
0.2	6.0	29%	14.7	16%	20	81
0.3	5.6	26%	11.0	13%	24	88
0.4	5.8	30%	7.2	8%	33	91
0.5	5.0	25%	6.4	8%	36	93
0.6	5.0	26%	4.4	5%	44	93
0.7	5.2	25%	3.0	4%	54	92
0.8	4.8	23%	2.7	3%	58	92
0.9	4.7	22%	2.4	3%	61	92
1	5.2	24%	0.9	1%	64	94
Scenario 2						
0.1	18.8	54%	24.1	22%	6	63
0.2	17.3	48%	17.4	17%	19	75
0.3	16.2	45%	14.8	15%	23	81
0.4	16.5	46%	8.2	8%	32	84
0.5	17.8	49%	5.2	5%	46	87
0.6	15.3	43%	4.3	5%	51	88
0.7	15.7	44%	4.0	4%	57	90
0.8	16.1	46%	3.5	4%	62	91
0.9	15.2	44%	3.4	4%	67	91
1	14.7	43%	2.7	3%	71	92
Scenario 3						
0.1	21.8	48%	19.7	19%	14	58
0.2	23.6	50%	18.9	18%	24	72
0.3	22.0	48%	16.8	17%	37	81
0.4	21.8	49%	13.2	13%	45	84
0.5	22.8	49%	12.2	12%	57	87
0.6	24.0	50%	10.9	11%	63	88
0.7	23.6	49%	8.1	8%	67	88
0.8	23.7	48%	7.8	8%	72	90
0.9	24.0	47%	6.0	6%	75	92
1	23.7	47%	5.7	6%	76	94

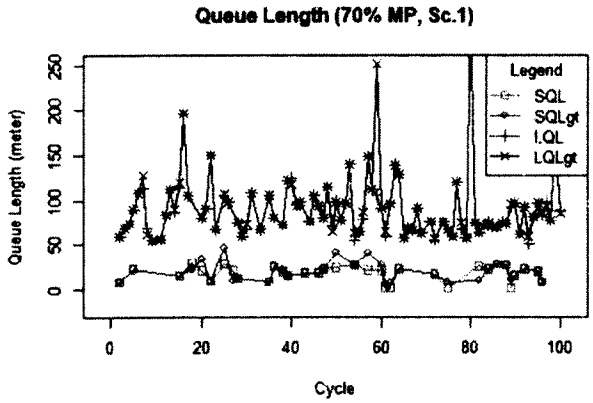
APPENDIX O - QUEUE LENGTH ESTIMATION BY SELECTING THE LAST PROBE FIRST

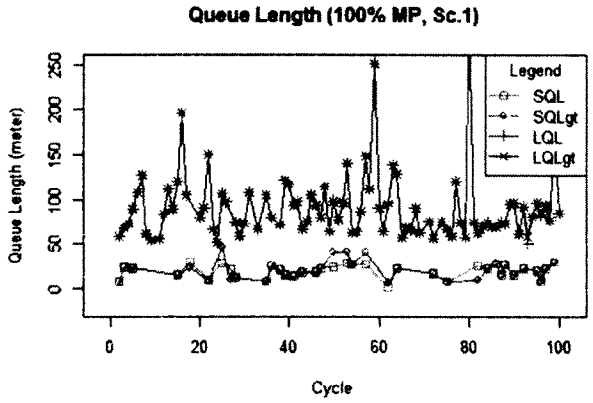
MP	MAE1	PerMAE1	MAE2	PerMAE2	ncycle1	ncycle2
Scenario 1						
0.1	7.7	35%	20.2	22%	11	62
0.2	6.4	30%	14.6	16%	17	80
0.3	4.9	23%	10.9	12%	21	87
0.4	4.7	24%	7.1	8%	26	88
0.5	3.9	19%	6.5	7%	25	89
0.6	4.1	22%	4.3	5%	30	89
0.7	3.9	20%	2.9	3%	35	88
0.8	3.4	18%	2.5	3%	36	88
0.9	3.0	15%	2.2	3%	37	89
1	2.7	13%	0.1	0%	37	90
Scenario 2						
0.1	19.1	59%	24.1	22%	4	62
0.2	12.0	43%	17.4	17%	10	75
0.3	12.9	41%	15.2	15%	13	78
0.4	10.9	40%	8.3	8%	15	80
0.5	7.5	30%	5.3	5%	17	83
0.6	3.4	15%	4.5	5%	15	83
0.7	4.7	20%	3.7	4%	16	84
0.8	5.4	24%	3.0	3%	18	84
0.9	3.8	16%	2.8	3%	18	82
1	3.5	15%	1.9	2%	18	80
Scenario 3						
0.1	21.9	47%	19.8	19%	13	51
0.2	20.6	44%	18.3	17%	17	61
0.3	18.1	41%	14.4	13%	19	68
0.4	10.7	27%	10.5	10%	16	71
0.5	14.7	34%	10.0	9%	20	67
0.6	18.1	39%	8.5	8%	21	67
0.7	18.1	38%	5.5	5%	23	70
0.8	17.3	37%	5.0	5%	23	73
0.9	13.8	30%	3.3	3%	21	74
1	11.0	24%	2.7	3%	20	75

APPENDIX P - QUEUE LENGTH ESTIMATION VERSUS GROUND TRUTH (SCENARIO 1)

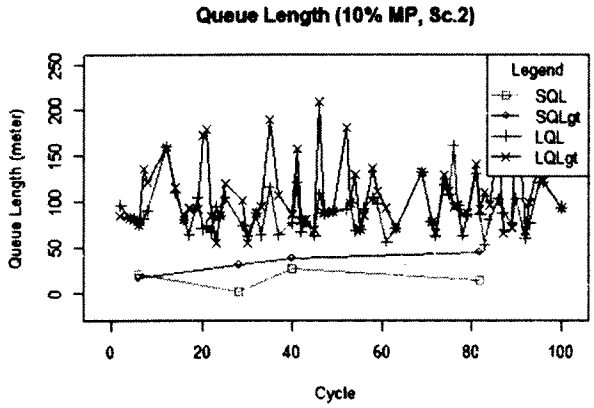


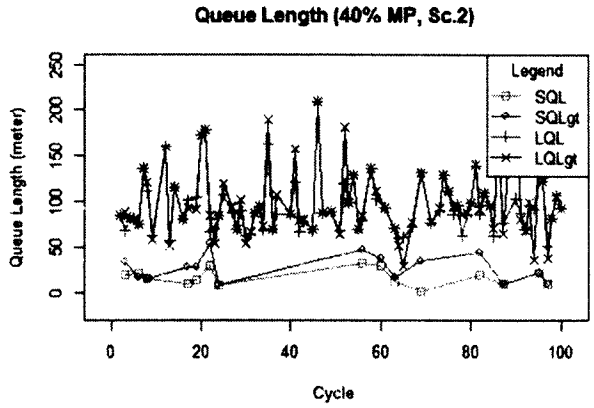


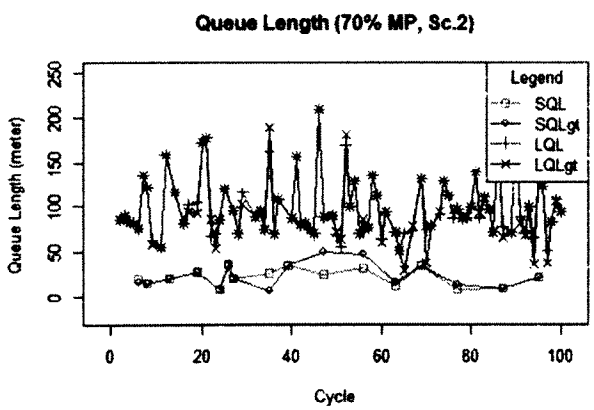


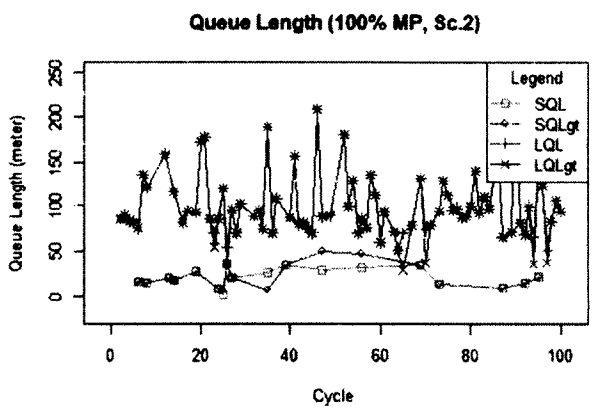


APPENDIX Q - QUEUE LENGTH ESTIMATION VERSUS GROUND TRUTH (SCENARIO 2)

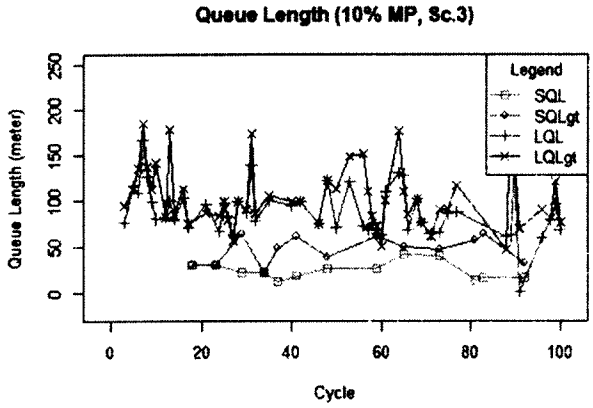


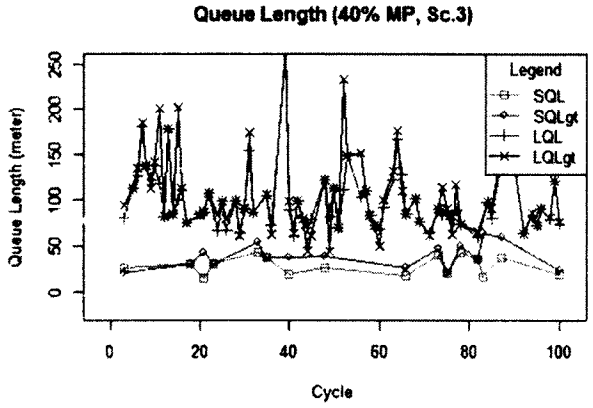


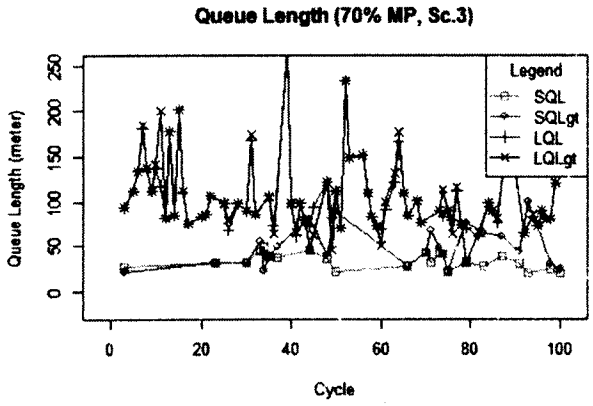


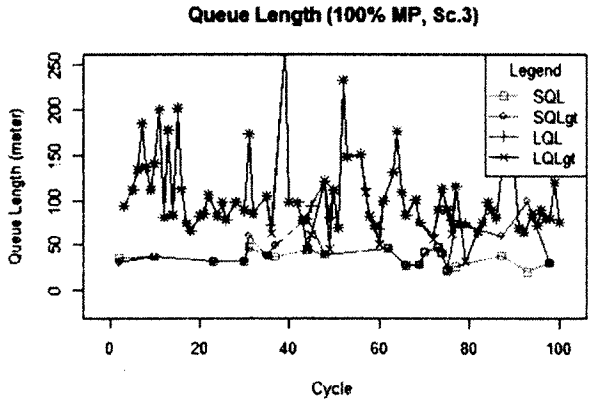


APPENDIX R - QUEUE LENGTH ESTIMATION VERSUS GROUND TRUTH (SCENARIO 3)









APPENDIX S - SUMMARY OF BEST REGRESSION MODEL

Deceleration - Logarithmic model

```
> summary(dec_logarithmic1)

Call:
lm(formula = deceldata1$SPEED ~ log(deceldata1$second))

Residuals:
    Min      1Q  Median      3Q     Max
-35.056  -6.944   1.146   7.209  45.839

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             48.35810    0.09742   496.4   <2e-16 ***
log(deceldata1$second) -10.62477    0.04501  -236.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.42 on 79506 degrees of freedom
Multiple R-squared:  0.4121,   Adjusted R-squared:  0.4121
F-statistic: 5.572e+04 on 1 and 79506 DF,  p-value: < 2.2e-16
```

Acceleration – Polynomial model

```
> summary(acc_polynomial1)

Call:
lm(formula = acceldata1$SPEED ~ acceldata1$second + I(acceldata1$second^2) +
    I(acceldata1$second^3))

Residuals:
    Min      1Q  Median      3Q     Max
-94.599  -4.373   2.100   5.549  53.986

Coefficients:
                        Estimate Std. Error  t value Pr(>|t|)
(Intercept)           -0.6832433  0.1261615   -5.416 6.14e-08 ***
acceldata1$second     10.9442236  0.0492897  222.039  < 2e-16 ***
I(acceldata1$second^2) -0.7220020  0.0049623 -145.498  < 2e-16 ***
I(acceldata1$second^3)  0.0118087  0.0001145  103.103  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.849 on 38142 degrees of freedom
Multiple R-squared:  0.6532,   Adjusted R-squared:  0.6532
F-statistic: 2.395e+04 on 3 and 38142 DF,  p-value: < 2.2e-16
```

APPENDIX T - FUEL CONSUMPTION MAE AND AVAILABLE CYCLES

p	Scenario 1				Scenario 2				Scenario 3			
	MAE		Cycle		MAE		Cycle		MAE		Cycle	
%	ltr/cycle	%	SQL	LQL	ltr/cycle	%	SQL	LQL	ltr/cycle	%	SQL	LQL
10	1.28	16%	11	62	1.83	17%	6	63	2.65	16%	14	58
20	1.00	14%	20	81	1.54	15%	19	75	2.24	17%	24	72
30	0.97	11%	24	88	1.38	15%	23	81	2.10	17%	37	81
40	0.78	11%	33	91	1.49	13%	32	84	2.15	14%	45	84
50	0.69	11%	36	93	1.48	13%	46	87	2.14	14%	57	87
60	0.64	11%	44	93	1.38	13%	51	88	2.00	14%	63	88
70	0.58	9%	54	92	1.31	12%	57	90	1.83	14%	67	88
80	0.50	9%	58	92	1.30	13%	62	91	1.78	14%	72	90
90	0.50	9%	61	92	1.15	13%	67	91	1.75	13%	75	92
100	0.65	10%	64	94	1.04	14%	71	92	1.66	14%	76	94

VITA

Semuel Yacob Recky Rompis was born in Manado, Indonesia, on August 15, 1976. He received a B.E. (*Indonesian : ST*) in Civil Engineering from Sam Ratulangi University in August, 2000 and M.Eng. (*Indonesian : MT*) in Civil Engineering in September, 2003 from the same university. In 2007 with the Australian Awards Scholarships, he studied Transport Systems Engineering at the Transport Systems Centre, University of South Australia in Adelaide. He received an M.Eng in Transport Systems Engineering from the university in 2009. He started his PhD study in Old Dominion University in 2011 with the Fulbright Scholarship program.