Summer 2016

Methodologies for Estimating Traffic Flow on Freeways Using Probe Vehicle Trajectory Data

Khairul Azfi Anuar
Old Dominion University

METHODOLOGIES FOR ESTIMATING TRAFFIC FLOW ON FREEWAYS USING PROBE VEHICLE TRAJECTORY DATA

by

Khairul Azfi Anuar
B.S. December 1998, University of Hartford
M.S. August 2008, Eastern Kentucky University
M.S. May 2012, Old Dominion University

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 2016

Approved by:

Mecit Cetin (Director)

Rajesh Paleti (Member)

Hong Yang (Member)

ABSTRACT

METHODOLOGIES FOR ESTIMATING TRAFFIC FLOW ON FREEWAYS USING PROBE VEHICLE TRAJECTORY DATA

Khairul Azfi Anuar
Old Dominion University, 2016
Director: Dr. Mecit Cetin

Probe vehicle data are increasingly becoming the primary source of traffic data. As probe vehicle data become more widespread, it is imperative that methods are developed so that traffic state estimators such as flow, density, and speed can be derived from such data. In this dissertation three different methodologies are proposed for predicting traffic flow or volume on a freeway. All of the proposed methodologies exploit several different traffic flow theories in conjunction with probe vehicle data to predict traffic flow. The first methodology takes advantage of the fundamental diagram or speed-flow relationship. The relationship states that flow can be estimated when speed is known. In this case, flow is traffic volume and speed comes from probe vehicles. Flow is predicted for four different models of fundamental diagrams and is analyzed at different time aggregation intervals. Results show that of the four fundamental diagrams, Van Aerde's Model is the best performing model with the lowest average percent error. It is also observed that flow prediction is more accurate during low speed (congestion) compared to high speed (free-flow) conditions. The second methodology exploits the shockwave theory, which pertains to the propagation of a change (discontinuity) in traffic flow. From probe vehicle trajectories, shockwave is estimated as the boundary between free-flow and congested regimes of traffic flow. After clustering the traffic regimes into free-flow and congested periods, the traffic flow during congestion is estimated using the Northwestern congested-regime fundamental diagram. From this estimation, the flow during free-flow is then predicted. Analyses

show that the percent error of the predicted flow during free-flow ranges from -9 to 1%. The third methodology is the car-following approach which relies on the spacing or distance between a leader and follower which can be directly measured from the trajectories. Based on a set of known probability distributions, the position of the follower vehicle with respect to the lead vehicle is estimated given that the spacing between the two random probe vehicles is known. A framework is developed to automatically process probe trajectories to extract relevant probe data under stop-and-go traffic conditions. The model is tested based on NGSIM datasets. The results show that when vehicle spacing is small the prediction of follower position is very accurate. As spacing increases the error in predicted follower position also increases. Though there exists some estimation error, all three approaches can reasonably predict flow for freeways using probe vehicle data.

This thesis is dedicated to my parents, Anuar and Shamsiah, whose upbringing values has mold me into the person I am today. I'm also dedicating this dissertation to my wife Wie and my son Keegan for their unconditional support.

ACKNOWLEDGMENTS

There are many people who have contributed to the successful completion of this dissertation. I am grateful to my supervisor, Dr. Mecit Cetin for his advice, guidance and support. It is an honor to work with him.

To my volleyball friends Alan, Jessica and Kathryn, you have provided me with a channel to relief my stress as things start to deviate from plan. Thank you. To my office partners, Filmon, Sem, SangHoon, Olcay and Elena, those late nights studies are finally paying off. Thank you for being there.

I am forever grateful to my parents, Shamsiah and Anuar for making me the person that I am today. Without their love and encouragement, I wouldn't know where I am today.

To my son Keegan, thank you for being patient with me. Now let's go find those Pokemon!

Finally, to my wife Wie, thank you for the sacrifices and being there for me. I would not have finished this dissertation without you.

NOMENCLATURE

q	Flow, vehicle per hour per lane
q_j	Flow during congestion, vehicle per hour per lane
$\hat{q}_j$	Estimated flow during congestion, vehicle per hour per lane
q_f	Flow during free-flow, vehicle per hour per lane
$\hat{q}_f$	Estimated flow during free-flow, vehicle per hour per lane
q_c	Flow at capacity, vehicle per hour per lane
u	Speed, miles per hour
u_B	Breakpoint speed, miles per hour
u_P	Probe vehicle speed, miles per hour
u_f	Free-flow speed, miles per hour
u_j	Speed during congestion, miles per hour
u_c	Speed at capacity, miles per hour
k	Density, vehicle per miles per lane
k_j	Jam density, vehicle per miles per lane
k_o	Optimum density, vehicle per miles per lane
η	Follower number
$\hat{\eta}$	Predicted follower number
X_s	Vehicle stop position, feet
X_g	Vehicle go position, feet
ΔX_s	Front bumper to front bumper spacing between two vehicles at stop position, feet
ΔX_g	Front bumper to front bumper spacing between two vehicles at go position, feet
g	Back bumper to front bumper spacing between two vehicles at go position, feet
L	Vehicle length, feet

$\bar{L}$	Average vehicle length, feet
w, ω	Shockwave speed, mile per hour or foot per second
$\widehat{w}, \widehat{\omega}$	Estimated shockwave speed, mile per hour or foot per second
F	Estimated value
O	Observed value
n, N	Number of samples
p	Market penetration rate of probe vehicles

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

CHAPTER 1

1 INTRODUCTION

Traffic state estimators such as speed and flow are important for a number of applications in traffic management and planning, such as signal timing, determining the number of lanes in highway design, etc. Traditionally, traffic data are collected from loop detectors. In recent years, another type of traffic data which are collected by probe vehicles (PV) are becoming increasingly popular.

PV are vehicles equipped with GPS-enabled devices such as cell phones or navigation systems. They typically record the location and speed of the vehicles themselves as they travel through the network. Since PV are mobile, they can collect traffic data all around the network. Due to their mobility, PV is increasingly becoming the source of traffic data. In comparison, loop detectors are installed at specific locations and can only provide information at fix points. To collect data for the network, multiple loop detectors must be installed. This then create a coverage and cost (installation and maintenance) issues for the transportation agencies who are responsible for the operations of the loop detectors. To tackle this issue, several studies have considered optimizing the locations of the loop detector within a transportation network (Bartin, Ozbay, & Iyigun, 2007; Fujito, Margiotta, Huang, & Perez, 2006; Kianfar & Edara, 2010).

In the current state, PV data are collected by third party organizations. The data come from fleet vehicles, taxis, vehicles with GPS-tracking smartphones, etc. These organizations collect the data and provide them to transportation agencies at a fee which is normally lower than the cost of installing and maintaining loop detectors. This makes PV data a cost efficient source of traffic data.

However, even with its associated cost, loop detectors provide data that are crucial to transportation agencies. Measuring all of the vehicles, loop detectors provide volume, speed and to an extent the density of traffic. These three traffic state estimators are commonly used to describe traffic conditions. PV on the other hand represent a fraction of the overall traffic as they collect speed when they travel through the network. It is common practice to assume that PV speed (sample) represents the speed of all vehicles (population). In general terms, PV is similar to loop detector in terms of recording traffic speed. Though they are slightly different speeds which are space mean speed (PV) and time mean speed (loop detector). But that's where the similarity ends because PV cannot directly measure flow or density.

As PV technology becomes popular, it has emerged as the potential primary source of traffic data. A study was performed (Kim & Coifman, 2014) to look into the feasibility of replacing loop detector with PV as traffic data for a state transportation agency. Therefore, as PV increasingly becoming popular, it is imperative that traffic state estimators other than speed and travel time can be measured or predicted from PV data. This leads to the main goal of this dissertation which is to predict traffic volume from PV data. In this dissertation, the term traffic volume will interchangeably refer to as flow or flow rate. In the end, they all point to the direction which is the number of vehicles.

Traffic flow is just one of the many traffic state estimators. Some other traffic state estimators that have been predicted from PV data are queue length (Anderson, Ran, Jin, Qin, & Cheng, 2011; Ban, Hao, & Sun, 2011; Cetin, 2012; Comert & Cetin, 2011) and trip origin-destination (Antoniou, Ben-Akiva, & Koutsopoulos, 2004; Barceló, Montero, Marqués, & Carmona, 2010; Caceres, Wideberg, & Benitez, 2007; Calabrese, Di Lorenzo, Liu, & Ratti, 2011)

PV data have also been used for density prediction (Juan C Herrera & Bayen, 2010; Hiribarren & Herrera, 2014; Roncoli, Bekiaris-Liberis, & Papageorgiou, 2015; Work et al., 2008) along with travel time prediction (Bar-Gera, 2007; Barceló et al., 2010; Chen & Chien, 2001).

While there have been several studies utilizing PV data, the main challenge when working with PV data is the penetration rate, which is the percentage of PV compared to the overall vehicles. Several studies have considered the challenges of penetration rate (Bucknell & Herrera, 2014; Chen & Chien, 2000; Cohen, Bosseboeuf, & Schwab, 2002; Patire, Wright, Prodhomme, & Bayen, 2015; Srinivasan & Jovanis, 1996). In the end most studies concluded that even at 5% penetration rate PV data are still reliable for making predictions on traffic states estimators. In the future, as the technology becomes more accessible it is expected that the penetration rate to increase.

The objective of this dissertation is to estimate volume of traffic from PV data specifically for freeways. Three different methods or approaches are proposed in this dissertation. All three methods exploit several different traffic flow theories in combination with PV data. Different types of datasets are applied to evaluate each method.

The first approach is called the fundamental diagram (FD) approach. A FD provides a relationship among the macroscopic traffic parameters, namely: speed (u), flow (q) and density (k). This approach relies mostly on the $u - q$ relationship. Given a robust and well-calibrated FD corresponding to the freeway of interest, one can estimate q corresponding to the u obtained from PV. This approach heavily depends on the goodness of fit of FD to the traffic data and the aggregation interval of the PV data.

The FD approach focuses on comparing the estimates of q obtained from a variety of single-regime FDs and aggregation intervals of the PV data. The FDs considered in this study are: Greenshields (1935), Underwood (1961), Northwestern (1967) and M. Van Aerde (1995). The aggregation intervals of the PV data considered are 5 minutes, 10 minutes, and 15 minutes.

The second method is called the shockwave (SW) approach. Shockwave w is the propagation of a change (discontinuity) in traffic flow. It was first proposed by the Lighthill-Whitham-Richards (LWR) model (Lighthill & Whitham, 1955; Richards, 1956). w refers to the boundary between free-flowing and congested traffic flow.

In the SW approach, *k*-means clustering is applied to differentiate the free-flow and congested regions. The resulting boundary between the two regions is w. From PV data congested speed u_j and free-flow speed u_f can be determined. The Northwestern congested regime FD is utilized to estimate flow during congestion $\hat{q}_j$. After completing the steps above, w, u_j, u_f and $\hat{q}_j$ are now known values. Using these four values, the flow during free-flow $\hat{q}_f$ can be estimated.

The third and final method is called the car-following (CF) approach. The main concept of this approach is exploiting the space headway between lead and follower vehicles. From vehicle trajectories, a specific pattern which is called a “ladder” can be observed. When a vehicle is in stop and go traffic, the stop motion is represented as a horizontal line within a vehicle trajectory. When these horizontal lines are combined for the different trajectories formed by multiple vehicles, they create a “ladder” shape within the trajectory. The steps within the “ladder” are the space headway between each vehicle.

In the CF approach, the term follower number η is commonly used. η is referred to the position of the follower behind a leader. As an example, if there are two vehicles and the first vehicle is in front of the second vehicle, then the second vehicle is $\eta = 1$. Now a third vehicle joins the group. This vehicle is $\eta = 2$ with respect to the first vehicle and $\eta = 1$ to the second vehicle.

From these "ladders" the distributions of the space headway are developed for each η. Given a space headway, the model would estimate η from the known distributions. η is predicted based on the distribution with the highest value utilizing the arguments of maxima.

Traffic state can be estimated by different variables. The most common and widely used traffic estimation from PV data is speed. For this application, it is inferred that PV speed (sample) represents the speed of general traffic (population). Travel time which is the inverse of speed can also be estimated from PV data using the same inference.

Other traffic state estimations such as density cannot be directly inferred from PV data since not all vehicles in the traffic stream serve as probes. Density is the number of vehicles occupying a space. In current practice, density is not an actual measurement but instead an estimation based on the ratio of flow over speed collected at a specific point. On the other hand, flow and speed is a direct measurement of the traffic.

Flow along with other variables such as density and travel time are important traffic state estimators. There are several applications where volume is a critical input. In transportation management and planning, volume is required to determine the number of lanes in freeway design. In traffic control, signal timing is dependent upon the flow at the intersection. In travel demand models, calibration of the model is dependent upon flow. If PV data ever to replace loop

detector data, it becomes imperative that methods are developed to estimate flow from PV data. Due to these practical aspects, flow is taken as the main parameter to be estimated. However, the methods presented here are equally applicable to density estimation.

All the approaches proposed in this dissertation predict the volume of traffic. In the FD approach, flow is predicted for any given PV speed, regardless of traffic condition. The SW approach identifies the traffic flow into free-flow and congested regimes. It then estimates volume for each regime. The CF approach predicts flow during congestion only where stop and go condition exists. The FD, SW and CF models are tested on Mobile Century (Juan C. Herrera et al., 2010), simulation and NGSIM data, respectively.

The contributions of this dissertation are:

- Improve methodology in predicting flow from PV and fundamental diagram by investigating four different FD models and different aggregation intervals.
- Implementation of *k*-mean clustering to classify free-flow and congested traffic flow from PV data. The resulting boundary between the free-flow and congestion is the shockwave.
- Predicting traffic flow for a congested region using Northwestern congested regime fundamental diagram. The prediction is based solely on PV speed.
- Propose a new methodology to predict flow by exploiting the shockwave which is observable from PV trajectory.
- Develop an algorithm to identify the stop and go movements of vehicles from their time-space coordinates.

- Predicting the number of vehicles between two PV by utilizing the stop and go movements of the PV.
- Creating links between the PV as the stop and go movements are located in multiple ladders. The links would reduce the overall time-space region to make a prediction on the number of vehicles.

This dissertation is proposing novel approaches to estimating flow from PV data in combination with several different traffic flow theories. Each approach is targeting different traffic flow conditions. While there have been studies on predicting traffic conditions from PV, the methodologies proposed in this dissertation are new and have never been investigated by other literatures.

The remainder of the study is organized as follows. Following this introductory section, review of literature is presented. This is followed by discussion on the methodological approaches and the data used for the analyses. Finally, the results are discussed, conclusions are drawn, and insights on future works are presented.

CHAPTER 2

2 BACKGROUND

Predicting traffic state estimators such as travel time, flow, shockwave and queue length have always been an important research topic in the field of transportation. This is because traffic estimators are used by traffic operation and management agencies to understand current traffic conditions and/or to forecast future conditions. As an example, the formation and dissipation of a shockwave provides clues on the build up, the length and the dissipation of a queue. Knowledge of the traffic state estimators can then be used for real time actions (e.g. intelligent transportation system) or long term solutions (e.g. transportation planning).

Regardless of the time periods (short or long term), accurate and reliable predictions of the traffic state estimators are crucial. These predictions are one of the several factors considered when key decisions are being made. Predictions of traffic state estimators are based on data collected from the roadways.

Traffic data are often collected by loop detector or similar devices (e.g. radar, microwave). They are called static sensors because they are installed at specific locations which then collect data for that specific location only. To collect data for the whole network, a vast amount of static sensors must be deployed throughout the network. They are normally deployed at specific intervals (e.g. every 0.5 or 1 mile) which may not be cost effective.

For a more scientific approach to the deployment of static sensors, Bartin et al. (2007) proposed a clustering method to determine the location of deployment of the sensors with a goal of obtaining travel time estimations with the fewest errors. In their approach, a sample roadway

is divided into cells. When vehicles travel through the roadway, the travel time error for each cell is calculated. Then the cells with the least error will be selected as the deployment locations of the static sensor. Instead of travel time, Kianfar and Edara (2010) utilized traffic speed to determine the location of a static sensor. Hierarchical and *k*-means clustering techniques were used to determine the location which were based on traffic speed collected from probe vehicles. In the city of Cincinnati and Atlanta, statics sensors are deployed by 0.5 mile (average) and 0.3 mile (actual), respectively. To determine the proper location of static sensors for both cities, Fujito et al. (2006) would systematically remove the sensors and compute the travel time index as a measure of performance. The configuration that resulted in the best performance measure was selected as the optimal locations for the sensors.

In addition to static sensors, traffic data are also being collected by PV (also referred to as floating car). PV are vehicles that are equipped with GPS tracking devices such as cell phone or GPS navigation. They collect data as they travel through the network. Because of their mobility to move freely around the network, PV are known as mobile sensors. Having one PV within a network would not provide sufficient data to predict traffic state estimators. It takes several PV (if not hundreds) to provide traffic data representing the entire system.

To help understand the difference between mobile and static sensors, a definition of both sensors is explained in terms of fluid motion. Mobile sensor or PV can be visualized as water traveling down a river. This is called the Lagrangian measurement. Meanwhile static sensors can be visualized as sitting on the river bank watching the water passing by which is called the Eulerian measurement.

Static sensors have been around longer than mobile sensors. Not surprisingly, transportation research utilizing mobile sensor data are enormous. But as mobile sensors becoming increasingly popular, research utilizing data from mobile sensors are also growing. Static sensors are commonly used to validate a traffic flow theory or to predict traffic state estimators. As mobile sensors become popular, the data that they collect are increasingly being applied to different aspects of transportation research. In areas where both datasets overlap, studies have also been performed by fusing them together.

In terms of traffic flow theory, Gordon F Newell (1993a) proposed a cumulative curve method for plotting cumulative number of vehicles to pass some location x by time t. By shifting the cumulative curve by the free-flow travel time between x_1 (arrival) and x_2 (departure), the horizontal displacement between x_1 and x_2 cumulative curves are the delay while the vertical displacement is the number of vehicles. Over the years, the name of this methodology evolved to what is known as the N-curve.

To validate the N-curve method, M. J. Cassidy and Bertini (1999) analyzed static sensor data collected near Toronto, Canada. In their study, they first used the data to plot the N-curve. From the plot they found that the predictions from the N-curve are comparable to nearby static sensor data. Using data collected from I-5 near San Diego, California, Kurada, Öğüt, and Banks (2007) performed similar studies on the N-curve. They've acknowledged that the N-curve is capable of explaining traffic flow phenomena while pointing out potentials for error due to bias in the data. Other studies have since follow suit using data collected from static sensors at different locations (Bertini & Cassidy, 2002; M. Cassidy & Mauch, 2001).

Navarro and Herrera (2014) proposed a new method to plot the N-curve from mobile sensor data. To validate their proposal, the R78 – Autopista del Sol which connects Santiago, Chile to the Pacific coast was selected as the study site. The departure location (bottleneck) is a toll plaza and the arrival location is 15 km upstream. The departure curve or the capacity is a line with a constant slope of 2,230 vehicle per hour (vph). The capacity was determined from static sensor located at the toll plaza. Arrival curve was then plotted using mobile sensor data. The arrival curve was shifted based on free-flow travel time between the arrival and departure locations. After plotting both arrival and departure curves, the study concluded that the predictions from these curves are comparable to the data collected at the toll booth.

Greenshields (1935) proposed the first fundamental diagram which assumed a linear $u - k$. Since then, several other models have been proposed by different researchers (Drake et al., 1967; Edie, 1961; Greenberg, 1959; Gordon Frank Newell, 1961; Underwood, 1961; M. Van Aerde, 1995; Wang, Li, Chen, & Ni, 2011). While the models differ from each other, the common denominator is that all of these models were developed based on static sensor data. In a new study, Li, Jian, and Monteil (2016) proposed a method to develop the FD from mobile sensor data. In their proposal, flow, speed and density were determined from mobile sensor data which were then translated into variables for the FD. Since their research was presented as a poster at the Transportation Research Board conference, no additional details can be found. Nonetheless, their research is mentioned in this study to highlight the growing field of research utilizing mobile sensor data to be applied to traffic flow theories.

Static sensors typically record the speed, flow and occupancy of traffic aggregated at specific time intervals for specific locations. While speed, flow and occupancy are important traffic estimators and are useful to transportation agencies, there are other estimators that are as

important but could not be collected by static sensors, such as travel time. From the road user standpoint, travel time is more important than traffic flow or speed.

Mobile sensors record their positon and speed as they travel along the roadways and therefor collect travel time data, which is the inverse of speed. While travel time is not explicitly broadcast, several providers are already providing traffic speed to the public, as illustrated in Figure 1. The public would then predict the impact of the speed to their expected travel time.

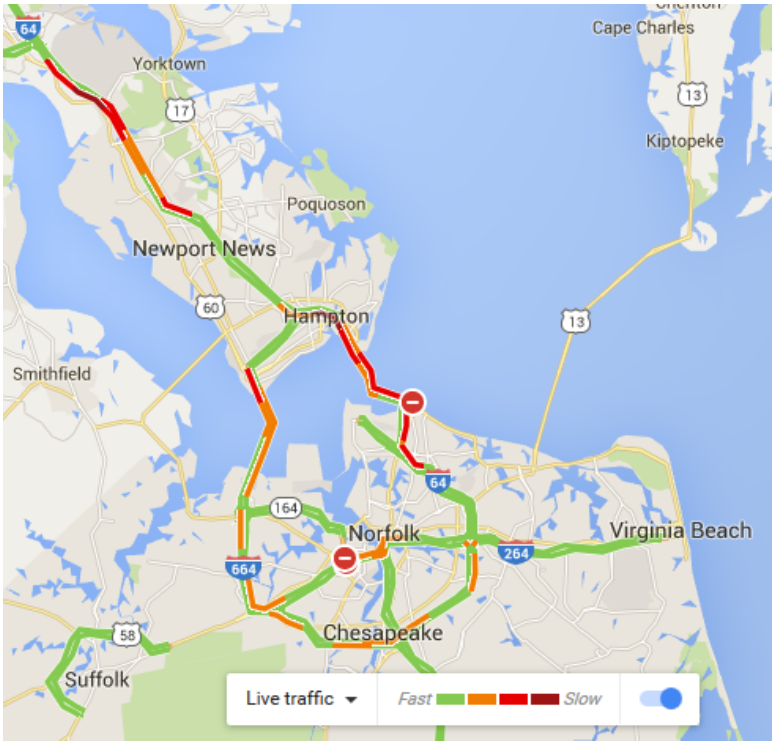

Figure 1 Traffic speed for the Hampton Roads region

In figure above, the traffic speed is for the Hampton Roads region of Virginia accessed through Google Maps on July 15, 2016 at 3:00 pm. Green indicates free-flow conditions and red indicates congestion. Incidents are highlighted with red circles.

It can be noted that static and mobile sensors complement each other instead of overriding one another. They each have different aspects of traffic data that they can collect. Due to the limitations as to what type of traffic estimators can directly be processed from the data, research have been on going into expanding the applicability of the static and mobile sensor data.

Static sensors record flow and speed of traffic at fix locations and therefore could not measure travel time. One approach to predicting travel time from static sensor data is the vehicle re-identification problem. The goal of this problem is to match a vehicle detected at one location with the same vehicle at another location. Travel time is then the difference between the times where the vehicles are detected at the different locations. As vehicles pass through static sensors, they would leave behind magnetic signatures. The task is then to match these match magnetic signatures as proposed by C. Sun, Ritchie, Tsai, and Jayakrishnan (1999) and Kwong, Kavaler, Rajagopal, and Varaiya (2009). During the vehicle matching process, errors occur where vehicles are mis-identified, identified more than once or not identified at all. To improve the vehicle matching prediction, Cetin and Nichols (2009) developed a two stage algorithm. The first step is to match similar vehicles using Bayesian method followed by solving an assignment problem. Coifman (2002) proposed another approach to predicting travel time from static sensor data by combining them with the LWR model (Lighthill & Whitham, 1955; Richards, 1956) and Gordon F Newell (1993b) fundamental diagram.

Similarly, mobile sensor data have been investigated for their applicability in predicting the speed, flow and density of traffic. Nanthawichit, Nakatsuji, and Suzuki (2003) used such approach by combining mobile and static sensor data and applying them to Payne (1971) macroscopic model and the Kalman filter technique. In their two study, Work et al. (2008) developed a partial differential equation (PDE) which was derived from the LWR model. The PDE was then applied to the cell transmission model (Daganzo, 1994) for the prediction of traffic flow and speed using PV data. Their work was extended by Juan C Herrera and Bayen (2010) to include a Newtonian relaxation and Kalman filtering to estimate traffic flow, density and speed.

Neumann, Touko Tcheumadjeu, Bohnke, Brockfield, and Bei (2013) proposed a methodology to estimate traffic flow from FD. Though their study may sound similar compared to the first approach of this dissertation, there are differences between them. First, they used a single FD for the same road type, e.g. same FD for all freeways. This fails to capture the spatial variations of traffic, e.g. difference in proportion of trucks and presence or absence of on-ramps. Moreover, they considered fitting only one type of FD with an aggregation interval of one hour. They extended their work by replacing the FD with Bayesian probability approach (Neumann, Touko Tcheumadjeu, & Bohnke, 2013). In this approach, flow is estimated from PV speed with the prediction relying upon Bayesian probability of historical data.

When video camera is mounted to a PV, the spacing between that vehicle and the vehicle in front of it can be calculated. From this information Seo, Kusakabe, and Asakura (2015a) proposed a methodology to estimate traffic flow, density and speed. Based on this idea, Seo, Kusakabe, and Asakura (2015b) incorporated additional data that are collected by PV such as speed and location to make predictions on traffic density.

All of the methods mentioned previously is a macroscopic approach in estimating volume. This means that the studies would predict volume for specific time-space regions. A microscopic approach would look into individual PV trajectory and make predictions from them. Considering a signalized intersection (Hao, Sun, Ban, Guo, & Ji, 2013) would predict the index of the PV. The indexing would consider both PV and non-PV. Their model was developed by using Bayesian network. After applying to NGSIM data, the results show the mean absolute error ranging from 0 to about 7.5% when the penetration rate is at 5%. The prediction improved as the penetration rate increased. By implementing variational formulation (Daganzo, 2005a, 2005b), the study by Z. Sun and Ban (2013) reconstructed the trajectory of non-PV at signalized intersection, given that PV trajectory are known. The results were measured in terms of error of number of vehicles within a queue. The error ranged from 1.2 to 2.5 vehicles with penetration rate varying from 20 to 100%.

PV data have also been used in queue length estimation (Anderson et al., 2011; Ban et al., 2011; Cetin, 2012; Comert & Cetin, 2009). For these studies the estimation of queue length are based on the propagation of the shockwave as PV approaches an intersection. Cai, Wang, Zheng, Wu, and Wang (2014) relied on loop detector data in addition to PV data to estimate queue length. Other studies have applied PV data on different topics such as transportation resiliency (Donovan & Work, 2015) and incident detection (Sethi, Bhandari, Koppelman, & Schofer, 1995).

Just like other types of data, there are challenges with PV data. When PV technology was first introduced, the penetration rate of PV is relatively low. Even at low penetration rates, studies were being performed in applying PV data to predicting traffic state estimators. As a result, several studies have been done to determine the acceptable PV penetration rate (Chen &

Chien, 2000; Cohen et al., 2002; Comert & Cetin, 2009; Patire et al., 2015; Srinivasan & Jovanis, 1996). Majority of the studies concluded that 5% penetration rate is acceptable for traffic analyses. As PV technology improves and becomes widespread, it is now being considered as the main source of traffic data replacing loop detector data (Kim & Coifman, 2014).

One of the methods proposed in this dissertation utilizes the knowledge of space headway between vehicles. The proposed method is formulated such that it requires the summation of distribution of the space headway. If the distribution is normal, then the sum of two or more normal distributions is also normal. Not so for lognormal distribution. Currently, there is no closed form for the sum of lognormal distributions. Several approximation methods have been proposed in the literature (Beaulieu & Rajwani, 2004; Beaulieu & Xie, 2004; Cobb, Rumi, & Salmerón, 2012; Lam & Le-Ngoc, 2006; Mehta, Wu, Molisch, & Zhang, 2007; Romeo, Da Costa, & Bardou, 2003; Schwartz & Yeh, 1982; Slimane, 2001; Zhao & Ding, 2007). All of them are complex and requires calculations of additional parameters to get the sum of the lognormal distribution. One of the oldest and simplest approximation is called the Fenton-Wilkinson (F-W) method (1960). It is commonly referred to by the other approximation methods (see references above). Due to its simplicity the F-W approximation is being investigated in one of the proposed methodology.

In this dissertation, three different methodologies to estimate traffic volume is presented. These methodologies take advantage of several traffic flow theories combined with PV data to predicting traffic volume. The following chapter is a discussion on the Methodology.

CHAPTER 3

3 METHODOLOGY

Three different approaches are proposed to estimate traffic flow on a freeway using PV data. Each approach will utilize PV data in combination with a traffic flow theory. The first approach will estimate traffic flow using PV combine with the FD. The second approach utilizes the shockwave w which is the boundary between free-flow and congested regions. The third approach utilizes the car following concept to estimate flow. The following subsections explain all of the proposed approaches beginning with the FD approach, followed by the shockwave approach concluding with the car following approach.

3.1 FUNDAMENTAL DIAGRAM APPROACH

Traffic stream is often described in terms of speed u, flow q and density k The relationship of these three variables can be represented in three different diagrams better known as the fundamental diagram (FD). The three diagrams that make up the FD describe the $q-u$, $q-k$ and $u-k$ relationship. The proposed methodology utilizes the $q-u$ relationship, where q can be estimated when u is known. Assuming a linear $u-k$ relationship, Figure 2 illustrates the FD $u-q-k$ relationships. In Figure 2b when u is known (i.e., u is the PV speed), q can be estimated.

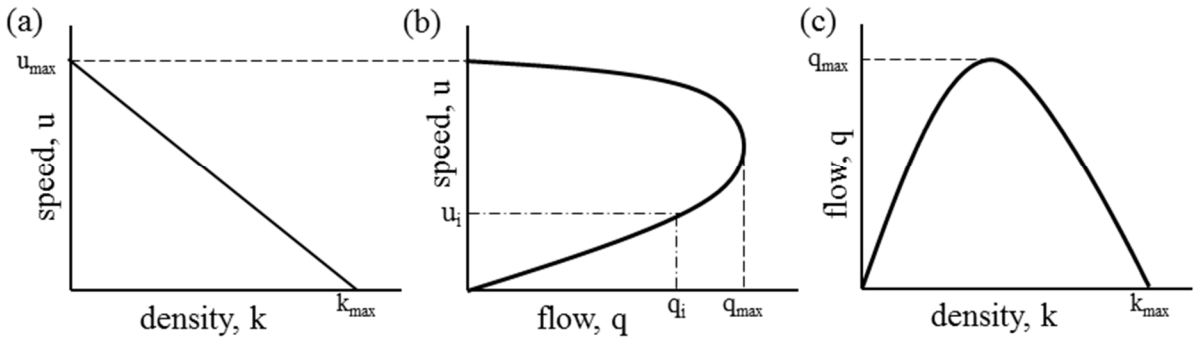

Figure 2 Fundamental diagram u- q- k relationship

Four different FDs are considered: Greenshields (1935), Underwood (1961), Northwestern (1967) and M. Van Aerde (1995). While there are several other FDs (Greenberg, 1959; Gordon Frank Newell, 1961; Wang et al., 2011), they are not considered in this study.

The first FD developed by Greenshields (1935) assumed a linear $u - k$ relationship that can be formulated by the equation:

$$k = k_j\left(1 - \frac{u}{u_f}\right) \quad (1)$$

where: k is density
k_j is jam density
u is speed
u_f is free-flow speed

In this relationship, as u approaches u_f then k approaches zero. As k approaches its maximum value k_j then u approaches zero. In real world, u_f is relatively easy to determine. This

is not the case with k_j where a typical value is between 185 and 250 vehicles per mile per lane based on vehicle length of 21 to 28 feet (May, 1990).

The second FD used in this dissertation was developed by Underwood (1961). This FD takes the form of a non-linear $u - k$ relationship that can be expressed as:

$$k = k_o \ln\left(\frac{u_f}{u}\right) \quad (2)$$

where: k is density
k_o is optimum density
u is speed
u_f is free-flow speed

For this FD, other than u_f, the other required known parameter is k_o which is the optimum density which occurs when q is maximum. k_o is difficult to observe and varies from one location to another (May, 1990).

A group of Northwestern University researchers (Drake et al., 1967) came up with an S-shaped curve to describe the $u - k$ relationship. Their proposed model was very similar to Underwood's model but with the introduction of a multiplier and a quadratic term. It can be formulated as:

$$k = k_o \left(2 \ln\frac{u_f}{u}\right)^{1/2} \quad (3)$$

where: k is density
k_o is optimum density
u is speed
u_f is free-flow speed

The final FD considered in this dissertation was developed by M. Van Aerde (1995) where k is formulated as the inverse of vehicle headway. To calculate the headway, four variables are introduced to the equation which can be formulated as:

$$k = \frac{1}{c_1 + \frac{c_2}{u_f - u} + c_3 u} \quad (4)$$

$$n = \frac{2u_c - u_f}{\left(u_f - u_c\right)^2} \quad (5)$$

$$c_2 = \frac{1}{k_j\left(n + {}^{1}/_{u_f}\right)} \quad (6)$$

$$c_1 = nc_2 \quad (7)$$

$$c_3 = \frac{-c_1 + \frac{u_c}{q_c} - \frac{c_2}{u_f - u_c}}{u_c} \quad (8)$$

where: k is density
k_j is jam density
u is speed
u_f is free-flow speed
u_c is speed at capacity
q_c is flow at capacity

Van Aerde FD works by performing a regression that would minimize the sum of squared errors (SSE) between the observed and predicted $u - q - k$ values in the orthogonal direction. A typical regression would calculate SSE by taking the difference either vertically (y direction) or horizontally (x direction) instead of orthogonally. To start the process, an initial starting point for u_f, u_c, q_c and k_j is selected. An iterative search is then performed until SSE is minimized. At the end of the search the values u_f, u_c, q_c and k_j are determined that would ultimately be used to calculate *k*.

Regardless of the different FD models, by utilizing the $q - u$ relationship, flow $\hat{q}$ can be estimated given a known u_P which is the PV speed. As the FD models vary from one another, for a given u_P, it is expected that $\hat{q}$ will also vary. This approach is then applied to data collected from the Mobile Century experiment (2010). The framework of the FD approach is illustrated in Figure 3.

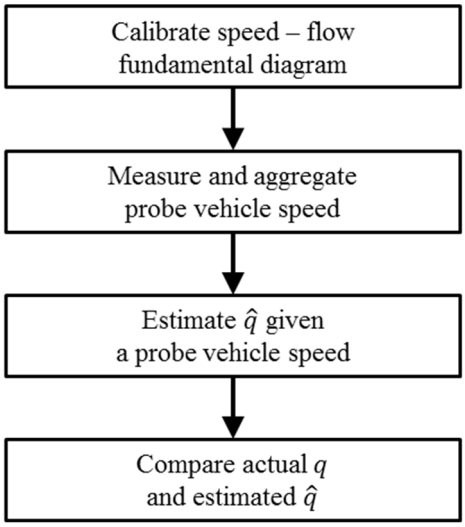

Figure 3 Framework for the fundamental diagram approach

The following subsection is a discussion on the shockwave approach, which is the second approach in estimating traffic flow.

3.2 SHOCKWAVE APPROACH

This methodology estimates traffic flow using PV data in combination with the shockwave theory as proposed by Lighthill-Whitham-Richards (LWR) model (Lighthill & Whitham, 1955; Richards, 1956). The LWR model is used to analyze traffic flow dynamics, in particular estimating the shockwave boundary and speed. Derived from an FD and the conservation law (Equation 9), the LWR model, also known as the kinematic wave model, describes the evolution of system state in terms of density, flow, or speed over time and space. The conservation equation can be formulated as:

$$\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = 0 \qquad 9$$

$\frac{\partial k}{\partial t}$ is the partial derivative of density k over time t and $\frac{\partial q}{\partial x}$ is the partial derivative of flow q over location x.

Shockwave w is the propagation of a change (discontinuity) in traffic flow. w refers to the boundary between free-flowing and congested traffic flow. It is considered backward (forward) moving as traffic transitions from free-flow (congested) to congested (free-flow). w can be formulated as:

$$w = {(q_j - q_f)}/{(k_j - k_f)} \quad (10)$$

Density k is the number of vehicles for a given length of a roadway at a specific time period. This is similar to taking an aerial photography of a roadway and counting the number of vehicles in that roadway section. In traffic flow theory, there is the fundamental relation of q as product of k and u. Rewriting this relation, k can be represented as:

$$k = {q}/{u} \quad (11)$$

Substituting 11 into 10:

$$w = \frac{q_j - q_f}{\frac{q_j}{u_j} - \frac{q_f}{u_f}} \quad (12)$$

Solving for q_f:

$$q_f = \frac{q_j - w\frac{q_j}{u_j}}{1 - \frac{w}{u_f}} \quad (13)$$

where: w is shockwave speed
q_j is flow during congestion
q_f is flow during free-flow
k_j is jam density
k_f is density during free-flow
u_j is speed during congestion
u_f is speed during free-flow

Relying solely on PV data, three variables - w, u_j and u_f - can be determined, leaving q_j and q_f as the two unknown variables. Figure 4 illustrates the relationship between vehicle trajectory and w along with the free-flow and congested regions. In this figure solid lines show the PV trajectory, dashed lines are non-PV trajectory and dotted line is w.

From PV data w is calculated by fitting a linear regression line at the boundary between free-flow and congested regions. From the regression w is the slope of the line. u_j and u_f are determined by the speed of PV in congested and free-flow regions. In this approach clustering is used to classify the two regions. The centers of the clusters or centroids are u_j and u_f. The clustering method is explained later in the chapter.

To solve the shockwave equation, one of the two unknown variables ($\hat{q}_j$ or $\hat{q}_f$) must be estimated. During free-flow period, traffic flow is considered erratic where there is fluctuation of vehicle flow and varying spacing between vehicles. In contrast, during congestion traffic flow is considered stable where vehicle flow and spacing between vehicles are relatively constant. Because of this characteristic, it is expected that by estimating $\hat{q}_j$ there would be less variation in $\hat{q}_f$ when solving the shockwave equation. Hence the decision to first solve $\hat{q}_j$ instead of $\hat{q}_f$.

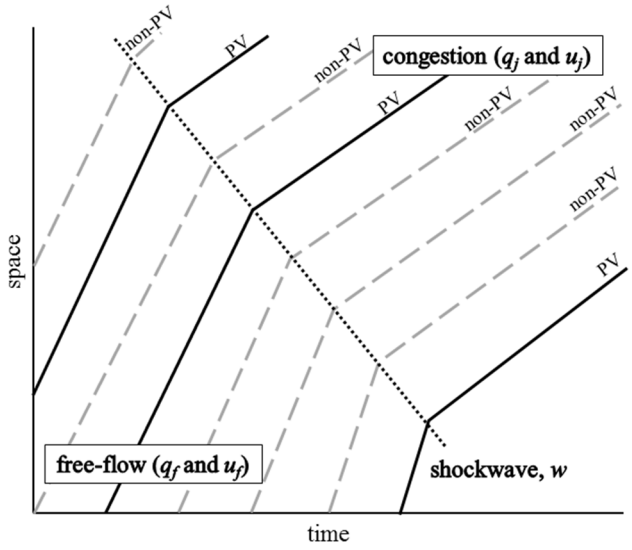

Figure 4 Example of vehicle trajectory with shockwave

To calculate $\hat{q}_j$ the Northwestern FD is re-visited. In the earlier section, the FDs are called single regime FD because they provide a single relationship for the entire traffic states (free-flow and congested). While a single regime FD would be applicable to all states of traffic, it has been determined that $\hat{q}$ has smaller errors during congestion compared to free-flow (Anuar, Habtemichael, & Cetin, 2015; Neumann, Touko Tcheumadjeu, Bohnke, et al., 2013). In other words $\hat{q}$ can be estimated by FD more precisely during congestion rather than free-flow period.

Instead of using the single regime FD a $u - k$ relationship specifically for the congested region proposed by the Northwestern group is implemented to calculate $\hat{q}_j$. The relationship can be formulated as:

$$u = u_B - 0.265k \quad (14)$$

where: u is speed
u_B is breakpoint speed (= 40)
k is density

To calculate $\hat{q}_j$,

$$\hat{q}_j = ku \quad (15)$$

Substituting 15 into 14:

$$\hat{q}_j = \left(\frac{u_B - u_P}{0.265}\right) u_P \quad (16)$$

where: $\hat{q}_j$ is estimated flow during congestion
u_B is breakpoint speed (= 40)
u_P is PV speed during congestion

After calculating $\hat{q}_j$ using Northwestern congested region FD and known w, u_j and u_f values from PV trajectory, $\hat{q}_f$ can now be solved. To calculate $\hat{q}_f$, plug in the known values into Equation 13. From simulated data, the resulting $\hat{q}_j$ and $\hat{q}_f$ are then compared to the observed q_j and q_f using several performance measures.

In summary, the framework of shockwave approach can be represented as a flow chart, shown in Figure 5.

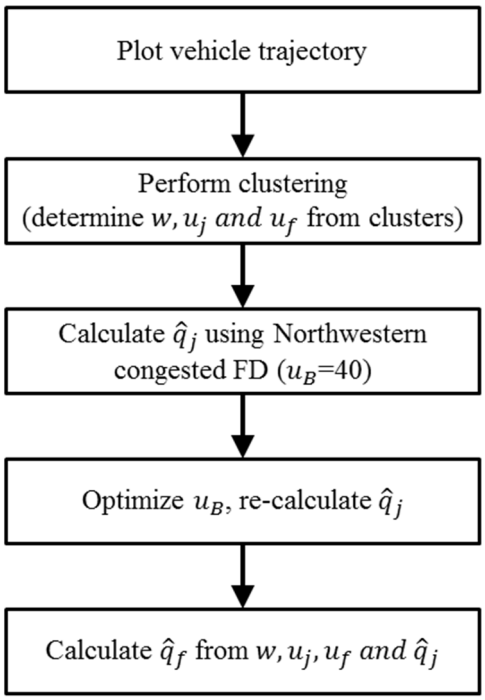

Figure 5 Framework for the shockwave approach

From the flow chart, there are several topics that have not been discussed. The clustering and the Northwestern congested FD are discussed immediately after this. While the optimization of u_B and the calculation of $\hat{q}_f$ will be discussed in the Results chapter.

3.2.1 Clustering

As mentioned earlier, clustering is applied to the data to identify free-flow and congested regions. Clustering is a technique where a set of data is partition into groups or clusters. It provides an understanding of how the data are naturally group. Clustering is normally performed on un-labeled data and is categorized as unsupervised learning. Using this technique members of a cluster are similar between them and are dissimilar with members from other clusters. It can be applied as a stand alone tool to provide intrinsic grouping of a set of un-labeled data or to be

applied as a step for other algorithms. A good clustering will produce clusters in which the similarity within the cluster (intra) is high while similarity between the clusters (inter) is low.

Clustering has been applied by practitioners and researchers alike. In real world applications, clustering has been applied to land use analysis to identify areas of similar land use. In marketing, it partitions the customers into groups which can be used to develop target marketing programs. It has also been used in pattern recognition, images processing and many other applications.

In the field of transportation research, clustering technique has been applied to determine the optimal location of loop detectors for improved travel time estimation (Bartin et al., 2007). Blasch, Banas, Paul, Bussjager, and Seetharaman (2012) used clustering technique to model a person's behavior (e.g. route choice). Clustering has also been used as a comparison against other techniques to group similar traffic patterns as demonstrated by Habtemichael, Cetin, and Anuar (2015).

Several clustering algorithms have been developed, each of which is based on its own unique principles. Due to the variations in the algorithm, the clustering technique can be divided into different methods. One proposal is to divide the clustering methods into two main groups: hierarchical and relocation (partitioning) methods (Fraley & Raftery, 1998).). Another proposal by Han, Pei, and Kamber (2011) divides the methods into four groups: hierarchical, partitioning, density-based and grid-based methods.

This study focuses on the partitioning method that works by moving objects from one cluster to another. To be specific, this study utilizes the *k*-means clustering. This is one of the most popular and simple clustering algorithm (Jain, 2010). It works by partitioning *n* number of

objects into *k* clusters in which each object belongs to the cluster with the nearest mean. The objective of *k*-means clustering is to minimize the total intra-cluster variance. The function to achieve this objective is formulated as:

$$Min\ Z = \sum_{i=1}^{k} \sum_{j=1}^{n} \left\| x_j - c_i \right\|^2 \quad (17)$$

where: x_j is observations
c_i is centroid
n is number of observations
k is number of clusters
$\left\| x_j - c_i \right\|$ is the Euclidean distance

Euclidean distance is the straight line distance between two points. While there are other distance measures such as Manhattan and Minkowski distances, the Euclidean distance are often used due to its simplicity. The steps for *k*-means algorithm is listed as:

1. Clusters the data into a pre-defined *k* groups
2. Select random centroids for each *k*
3. Assign objects to their closest centroids according to the Euclidean distance
4. Recalculate the new centroids
5. Recalculate the Euclidean distance between each object and new centroid
6. Repeat step 3 until no objects move between clusters

The attractiveness of *k*-means clustering includes: efficient and easy to understand and implement. The disadvantages are: results dependent upon initial centroid assignment, pre-

defined value of *k* and different distance error will give different results. In this study, the *k*-means clustering is implemented by using the built in function developed within the statistical software R.

3.2.2 Northwestern Congested Regime Fundamental Diagram

The Northwestern congested regime FD is used to estimate $\hat{q}_j$, which is the estimated hourly flow during congestion. It is then compared to q_j which is the observed hourly flow. The concept of Northwestern congested FD is shown in Figure 6. The solid line is the Northwestern congested FD and for reference purpose the dotted line is the linear single regime FD. As the name implies, the Northwestern congested FD is fitted to congestion data only. In comparison a single regime FD is fitted to both free-flow and congested data. By fitting to different types of data, it results in different slopes and intersect values for the different FDs.

In the Northwestern congested FD, the term $u_B = 40$ which is the breakpoint speed is introduced. This is the speed at which traffic flow transitions from free-flow to congestion and vice versa. When this transition points are connected together, they form $\hat{\omega}$. Note the difference between w and ω. While they both stands for shockwave, in this dissertation the term w is used for flow estimation while ω is used to determine u_B.

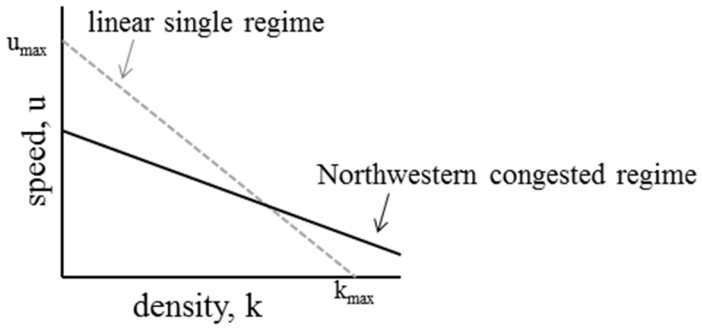

Figure 6 Northwestern congested regime fundamental diagram

By adjusting u_B it changes the transition points which in turns affects $\hat{\omega}$. For a given u_B, $\hat{\omega}$ is estimated by fitting a linear regression line through a group of transition points. As the time-space coordinates of these transition points changes, they also affect $\hat{\omega}$. In addition to the coordinate, another factor that affects $\hat{\omega}$ is the number of transition points.

In this approach, rather than calculating $\hat{\omega}$ for a large number of transition points, the transition points themselves are divided into groups. This would capture the different $\hat{\omega}$ as the vehicle demand changes. Each group is assigned a specific number n of PV. Too small of an n per group may cause severe fluctuation of $\hat{\omega}$ while too large of an n may result in averaging the $\hat{\omega}$ over too large of time period. As a result, from the total number of PV, they are divided into p groups containing n PV.

To determine the proper u_B the $\hat{\omega}$ for each group is compared against a known ω. The objective is to adjust u_B so that the difference between $\hat{\omega}$ and ω is minimized. The objective function can be formulated as:

$$Min\ Z = \sum_{i=1}^{p} \frac{|\widehat{\omega}_i - \omega|}{p} \quad (18)$$

where: $\widehat{\omega}_i$ is estimated shockwave for each group
ω_i is known shockwave for each group
p is total number of groups

After selecting u_B the variables - w, u_j and u_f - can be calculated. They are then plugged into Equation 13 to calculate $\hat{q}_f$. To measure the accuracy of the results, the difference between $\hat{q}_f$ and q_f are calculated. The following section is a discussion on the car-following approach.

3.3 CAR FOLLOWING APPROACH

Car following describes the behavior of a vehicle (follower) with respect to the vehicle in front of it (leader). In a traffic stream, the follower will adjust its behavior (such as speed and spacing) to safely trail the lead vehicle. As the leader decreases its speed, the follower will do the same while maintaining a safe spacing. During congestion, vehicles stop and go as a result of following the vehicles in front of them. As the lead vehicles stops or slows down, the followers adjust their vehicle speed. The car-following (CF) approach exploits the stop and go movements of vehicles. When the trajectories of these vehicles are plotted, the stop and go movements are represented as horizontal or near horizontal lines. When these lines are combined, they form a "ladder" shape, as shown in Figure 7.

In Figure 7, the lines are trajectories for vehicles 1, 2 and 3. In this example, vehicle 1 is travelling at a specific speed when it slows down and then speeds up again. This slow down movement can be observed by the change of slope of the vehicle trajectory. As a result of this

slow down, vehicle 2 and 3 encountered stop and go conditions as indicated by the horizontal lines of their trajectories. Together, the slow down of vehicle 1 combined with stop and go of vehicles 2 and 3 create a ladder shape trajectory. The CF approach presented here exploits the regular patterns in this ladder to predict the number of (unobserved) vehicles in between two probe vehicle trajectories.

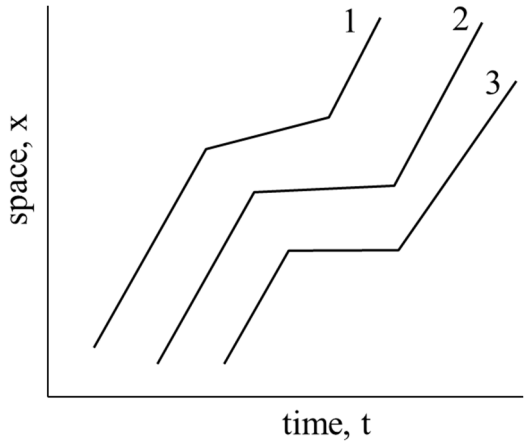

Figure 7 Hypothetical vehicle trajectory

After identifying the ladder the spacing between each leader and follower pair is calculated. In Figure 7 the leader and follower pairs are vehicles 1-2, 1-3 and 2-3. In this example vehicle pairs 1-2 and 2-3 are for follower number one ($\eta = 1$) where vehicle 2 is follower one compared to vehicle 1 and vehicle 3 is follower one compared to vehicle 2. Vehicle pair 1-3 is for follower two ($\eta = 2$) where vehicle 3 is the second follower with respect to vehicle 1. The pairs continue with additional vehicles 4, 5 so on and so forth.

For each follower number, there are multiple observations of spacing. Referring to Figure 7, there are two (1-2 and 2-3) and one (1-3) observations for $\eta = 1$ and $\eta = 2$, respectively. From these observations, the distribution of the spacing can be observed. To predict the follower number η for a given vehicle pair the spacing between the two vehicles is used. To predict η the naive Bayes model is implemented.

The Naive Bayes algorithm is based on conditional probabilities. It uses Bayes' theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. For a given spacing between two vehicles, the distribution with the highest probability will be chosen as the predicted η.

To find the ladder and the spacing from the PV trajectory, an algorithm is developed. The spacing which is an output of the algorithm is then analyzed for its distribution. This learning process or training would create a set of known distributions of the spacing for each η. Given a spacing also known as the testing stage, the model would then predict η. Performance measures are used to evaluate the accuracy of the prediction. The algorithm, training, testing and the performance measures are discussed next.

3.3.1 Algorithm

The algorithm for this approach is more complex compared to FD and shockwave approaches. Multiple steps with different calculations and assumptions are implemented as part of the algorithm. Figure 8 shows the framework of the algorithm.

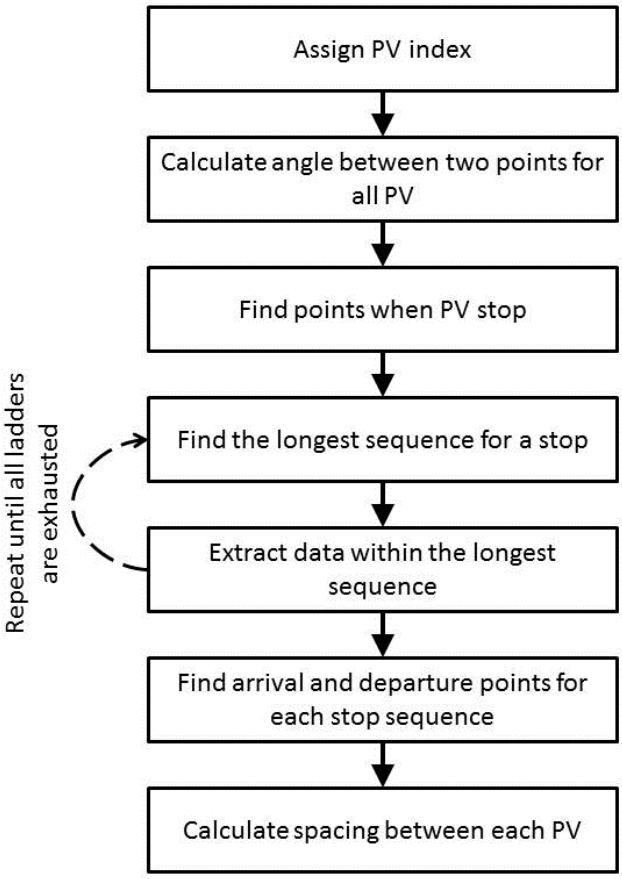

Figure 8 Framework for the algorithm of the car-following approach

Prior to discussions on the algorithm, the assumptions are:

a. ΔT = 5 seconds

b. w = -18 fps (similar to Lu and Skabardonis (2007) findings)

c. Minimum length of stop (l_s) = 3 seconds

d. Speed threshold (u_{lim}) = 0.1 fps

e. Maximum number of stop (s_{max}) = 5 (unit less)

The steps for the algorithm are listed as:

Step 1: Determine the vehicles sequence based on the times the vehicles exit the study area. The first vehicle exiting the area (based on time) is vehicle number one. Followed by vehicles two, three, etc. This vehicle sequence will be used as ground truth measurements.

Step 2: Calculate the angle between two consecutive points for all points for all vehicles, with exception of the last points. The angles are calculated by using the arc tangent trigonometry function, as shown in Figure 9.

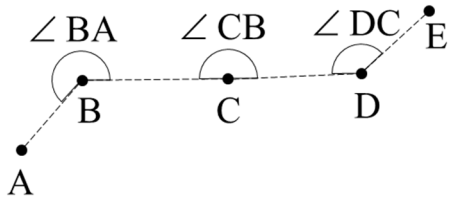

Figure 9 Angle between two points

The arc tangent is selected due to the structure of input data and the output is always positive. While other trigonometric functions are available to calculate the angle, the output could be negative values which make creates unnecessary conversions or calculations.

Step 3: Determine the points for each vehicle when their movements are stopped. In this algorithm, a stop is defined when the vehicle speed is less than or equal to u_{lim}.

Each stop for every vehicle is identified and label as stopID. The minimum number of observations *N* in each stopID is one, but there's no maximum limit.

Step 4: Find the stopID with the maximum *N* provided that *N* is greater than or equal to l_s. This stopID is identified as the longest stop sequence. Then determine the start and end points from this longest stop sequence. Extract data between the start and end points by imposing two lines with a slope of w extending out from the start and end points by ΔT. This extraction method is shown in Figure 10.

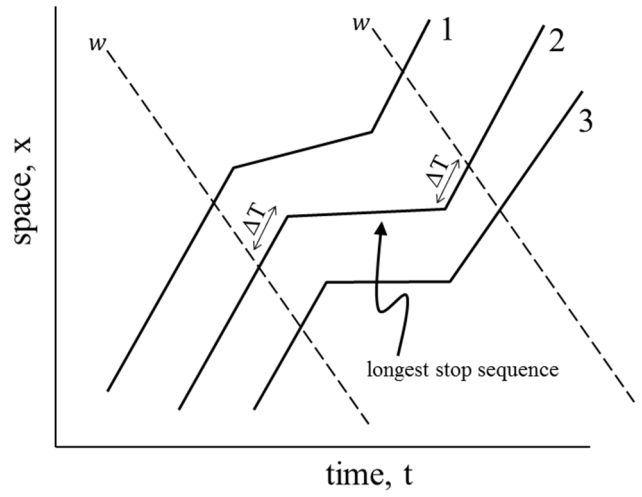

Figure 10 Extraction window

Remove the extracted data from the original dataset. Identify the extracted data as extractedID. Repeat this step until all observations are exhausted.

Step 5: For each extractedID find the arrival and departure points. The arrival point X_s is where the angle is at maximum while the departure point X_g is where the angle is

at minimum (see Step 2). The maximum number of extractedID allowed is less than or equal to s_{max}.

Step 6: Calculate the X_s and X_g difference between the leaders and all of their followers. Perform calculation until all leaders are exhausted. Figure 11 is an illustration on how to calculate the X_s differences, which are labeled as ΔX_s. Though not shown, the ΔX_g calculation is the same as ΔX_s.

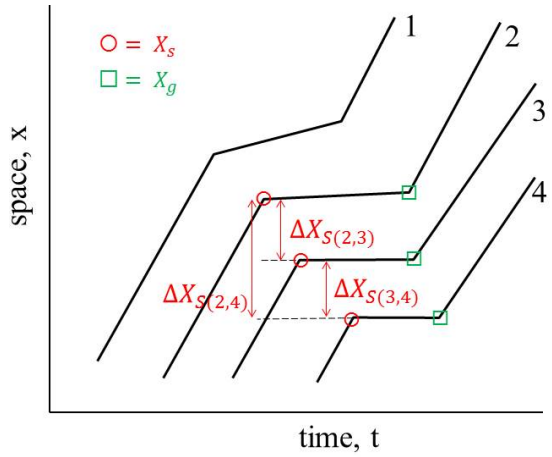

Figure 11 Locations of X_s, X_g and ΔX_s

In figure above, vehicle 1 is the leader for vehicles 2, 3 and 4. Vehicle 2 is then the leader for vehicles 3 and 4. From the viewpoint of vehicle 2, vehicle 3 is η = 1 while vehicle 4 is η = 2. Vehicle 3 is the leader for vehicle 4, which makes vehicle 4 η = 1 to vehicle 3. So on and so forth if there are additional vehicles.

ΔX_s and ΔX_g are calculated for each vehicle pair (e.g. 1-2, 2-3, 3-4, etc). Recall earlier that η is the follower number or position of the follower behind a leader.

3.3.2 Model training

To validate a model, different sets of data are applied. Typically, a training data is applied to train the model followed by test data to test the model. This type training and testing has been performed by a handful of studies (Hou, Edara, & Sun, 2012; Khodayari, Ghaffari, Kazemi, & Braunstingl, 2012; Wei & Liu, 2013).

To train the model, a set of training data is applied to the algorithm. In the training data, the information for all vehicles are known, including but not limited to positions, timestamps and vehicle lengths. The first step of the training is to identify the vehicle positions in terms of leader and follower. Then ΔX_s and ΔX_g are calculated for every pair of leaders and followers. Each leader-follower pair represents an η. Using Figure 11 as an example, vehicles 2-3 and 3-4 are the vehicle pairings for η = 1. Vehicles 2-4 is the vehicle pair for η = 2. If vehicle 5 exists and located behind vehicle 4, then vehicle 3-5 is the pair for η = 2 while vehicle 2-5 is the pair for η = 3. So on and so forth until all vehicle pairs are identified.

Regardless of the vehicle movement (stop or go), ΔX is the spacing between the front bumper of a vehicle to front bumper of another vehicle. ΔX can be expressed as the length of the vehicle L plus the gap g between back and front bumper. If there are more than two vehicles, then L is the sum of all the vehicle lengths and g is the sum of the vehicle gaps. This relationship can be formulated as:

$$\Delta X = \sum_{i=1}^{n-1} g_i + \sum_{i=1}^{n-1} L_i \quad (19)$$

Figure 12 is an example of the relationship between ΔX, L and g. Their relationship can be formulated as:

$$\Delta X = (g_1 + g_2 + g_3 + g_4) + (L_1 + L_2 + L_3 + L_4) \quad (20)$$

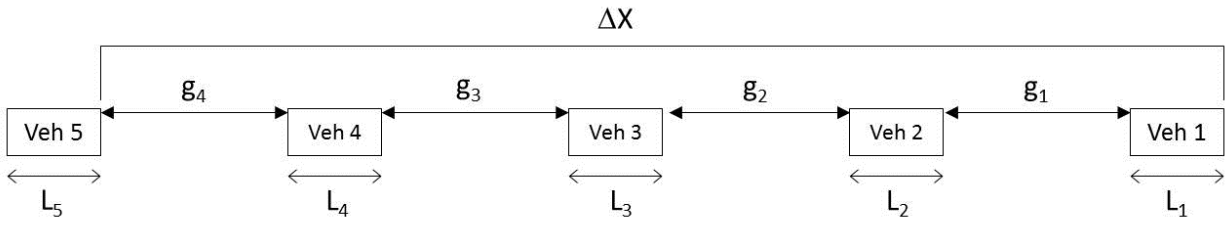

Figure 12 Example of vehicle spacing, gap and length

The purpose of defining ΔX is to differentiate the random and constant variables. If ΔX is used as the key variable, it should be noted that ΔX is combination of random variable g and constant L. For analyses, it is imperative that the variable to be random. Hence the decision to used random variable g as the primary variable to predict η. Since information for all vehicles are available in the training data, g can easily be calculated. To calculate g, simply deduct L from ΔX. After determining g, the pdf is calculated for every η. The crucial pdf parameters are the mean μ_η and variation σ_η^2. Figure 13 is an example of the pdf of g for η = 1 to 4. In the training

model, the pdf of g is calculated for every η. In addition to the variable g, the pdf of variable ΔX is also calculated for every η.

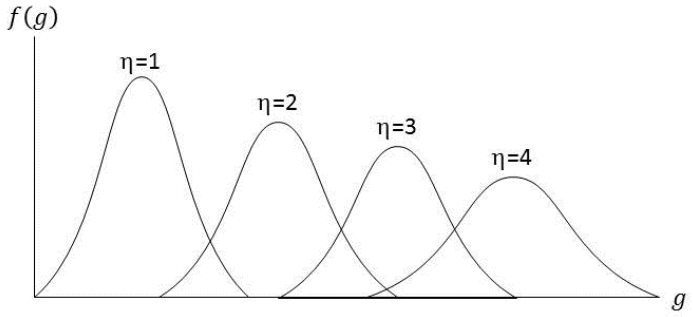

Figure 13 Hypothetical example pdfs of g

Random variable g is assumed to have either normal (21) or lognormal (22) distribution with mean μ and variance σ^2 if its pdf are:

$$f(g) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(g-\mu)^2}{2\sigma^2}\right] \quad 21$$

$$f(g) = \frac{1}{g\sigma\sqrt{2\pi}} exp\left[-\frac{(ln[g]-\mu)^2}{2\sigma^2}\right] \quad 22$$

It is imperative to point out that the prediction of η is highly dependent upon the pdf of g. The type of distribution and the variable used to determine the pdf impact the prediction of η. It is important that the data used to estimate the pdf come from a random sample. The pdf of g is

tested for types of distribution by using Cullen and Frey (1999) technique. This technique is a plot of skewness and kurtosis which provides some insight as to the type of distribution to fit the data. On this plot several common distributions are displayed. For normal, uniform, logistic and exponential distributions, there is only one possible value for skewness and kurtosis. They are represented as single points on the plot. For gamma and lognormal distributions, the possible values are represented as lines.

As stated earlier, the pdfs are calculated for the variable g. To ensure that g is random, the conditions of overlap and duplicate are removed from the training data. Overlap is a case where the vehicle pairings to calculate g intersect with one another. This would result in some correlation between the vehicle pairings. As an example, from Figure 12, when $\eta = 2$, the pairings are vehicles 1-3 (with vehicle 2 between), 2-4 (with vehicle 3 in between) and 3-5 (with vehicle 4 in between). The overlap occurs for vehicles 1-3 and 2-4 where vehicle 2 is used twice to calculate g. The overlap also occurs for vehicles 2-4 and 3-5 where vehicle 3 is used twice. To eliminate the overlap, only consecutive vehicle pairs are considered. Using the same figure as an example, for $\eta = 2$ the consecutive vehicle pairs would be 1-3 (with vehicle 2 in between) and 3-5 (with vheicle 4 in between). By selecting consecutive vehicle pairs, the issue of overlap is eliminated.

Another issue for calculating g is duplicate. This condition exists when vehicle pairings are located in multiple ladders. Using Figure 11 as an example, imagine two additional sets of ladders created from the continuation of the same trajectory, one to the left (upstream) and one to the right (downstream) of the existing ladder. All three sets of ladders have the same vehicle pairs. Even though the ladders are located at different time-space coordinates, the vehicle

pairings are still the same. To eliminate the duplicates, vehicle pairs can only be located in a single ladder.

3.3.3 Model testing

The model is then applied to test data, which is different than training data. The variables for the test are the same as training data. The only difference is that the actual η are hidden from the model. While the variable is there, it is not being used to estimate η. It is used only as a comparison against a prediction.

Although the test data consist of data for all vehicles, only specific vehicles are randomly selected to be assigned as PV. Testing would be performed on the PV data while the original test data consisting of all vehicles will be used as ground truth comparison. From the trajectory, ΔX_s and ΔX_g of the PV can be determined. In this method, only ΔX_g is considered. In real world ΔX_g is more stable compared to ΔX_s. When a vehicle is in a stop motion and is ready to move, it will do so immediately. ΔX_g is the point where the vehicle starts moving. However, when a vehicle is in motion and encounters a stop condition, the point where the vehicle stop ΔX_s is difficult to define. A vehicle might pull up to the stop location and brakes (instantaneous stop) or a vehicle might do a slow roll as it approaches the stop location. With this slow roll motion, it is difficult to identify ΔX_s because the vehicle is still moving albeit slowly. As it approaches the stop vehicle in front of it, it may not even stop completely if the front vehicle starts moving again. Because of this, it is more desirable to use ΔX_g for analyses. From here on ΔX_g will be referred to as ΔX.

To predict η the observed ΔX needs to be transformed to g so it could be applied to the pdf obtained from the training data. From PV data, the only variable that can be measured is ΔX. g

and L are unknown values. To solve for g, given that ΔX is known, L needs to be estimated. Where L is the product of number of vehicles S and the average length of vehicles $\bar{L}$ in front of the last follower. To estimate L, the pdf of ΔX from the training data is utilized.

The additional step of estimating L is required because the random variable g is not observable from PV data. Recall that $\Delta X = g + S\bar{L}$. The term $S\bar{L}$ varies as S increases ($\bar{L}$ is constant). To estimate S, find the argument of the maxima from $f(\Delta X)$ which can be formulated as:

$$\hat{S} = \underset{S}{\operatorname{argmax}} f(\Delta X|\eta) \qquad 23$$

Figure 14 illustrates the process of finding the argmax $f(\Delta X|\eta)$. The first step is to calculate L for S_{ini} = 1, …, N. For each L find the $f(\Delta X|\eta)$. Then find the maximum of $f(\Delta X|\eta)$ and assign η to S. In this figure, when S_{ini} = 2, the resulting argmax is S = 3.

Though it may look cumbersome, S needs to be determined so that ΔX can be transformed into g. The purpose of using g is to ensure that the variable is random. However, S introduces a new a variable and therefore a new uncertainty to predicting η.

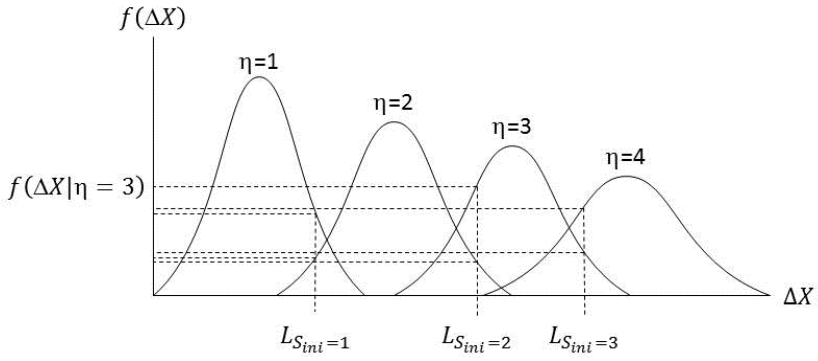

Figure 14 Example of argmax $f(\Delta X|\eta)$

After predicting S for a given ΔX, the variable g can be calculated. To predict η, the same approach of argument of maxima is utilized. The formula can be written as:

$$\hat{\eta} = \underset{\eta}{\text{argmax}}\, f(g|\eta) \qquad 24$$

Figure 15 illustrates the process of finding the argmax $f(g|\eta)$. Given a value g_i, the argmax will find the maximum value from each pdf of g for every η. The η with the maximum value will then be selected as the predicted η.

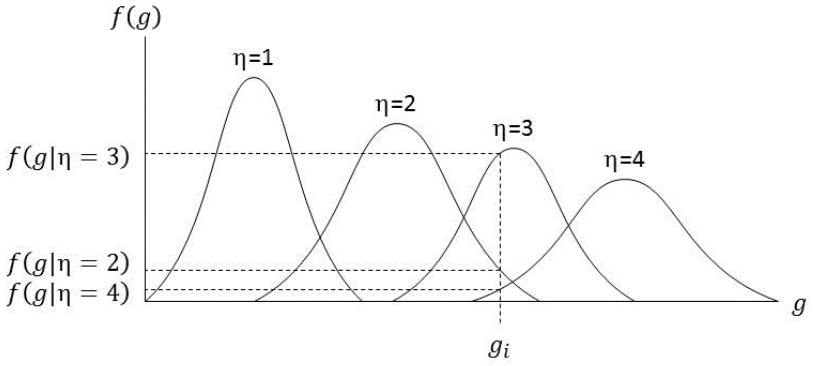

Figure 15 Example of argmax $f(g|\eta)$

As with any model, there are errors in prediction. As indicated earlier the higher the vehicle spacing, the greater the variations of g, resulting in increased prediction error. To reduce error, instead of making a direct prediction from leader to the last follower, the prediction will take a path utilizing intermediary vehicles in between the two vehicles, if they exist. Each path consists of links that connect the PV. Each link has an error cost ε associated with it. To predict η is to follow the path with the least amount of ε. Since the intermediary vehicles are actual PV, their g's are also known. Hence η can be predicted for these intermediary PV. As an example, refer to Figure 12. Let's assume that vehicles 1, 3 and 5 are PV1, 2 and 3, respectively. In this figure the intermediary PV is vehicle 3. To predict η from vehicle 1 (PV1) to vehicle 5 (PV3), a direct prediction can be made from PV1 to PV3 or a multiple links prediction can be made from PV1 to PV2 to PV3. The decision on which path to take is dependent upon the ε for each link. As stated earlier, the path with the least ε will be chosen. This is similar to solving the shortest problem.

3.3.4 Sum of distribution

Let random variables X and Y to be normally distributed, with $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$. The sum of both random variables $Z = X + Y$ will result in $Z \sim N(\mu_Z = \mu_X + \mu_Y, \sigma_Z^2 = \sigma_X^2 + \sigma_Y^2)$. This is true for the sum of any normal distribution.

Currently, there is no closed form for the sums of lognormal distribution. While there have been several approximation methods, the simpler F-W method (1960) is investigated. Consider the sum of M lognormal random variable X where each $X_m \sim LN(\mu_{X_m}, \sigma_{X_m}^2)$, the expected value and variance of X are $E(X) = M * E(X_m)$ and $Var(X) = M * Var(X_m)$. The F-W approximate such that:

$$exp\left(\mu_X + 0.5\sigma_X^2\right) = M * E(X_m) \tag{25}$$

$$[exp(\sigma_X^2) - 1] * exp(2\mu_X + \sigma_X^2) = M * Var(X_m) \tag{26}$$

Solving for the expected mean μ_X and variance σ_X^2:

$$\sigma_X^2 = \ln\left[1 + \frac{exp\left(\sigma_{g_m}^2\right) - 1}{M}\right] \tag{27}$$

$$\mu_X = \ln\left(M * \exp\left(\mu_{X_m}\right)\right) + 0.5(\sigma_{X_m}^2 - \sigma_X^2) \tag{28}$$

By solving for μ_X and σ_X^2 the ε for any η can now be approximated, assuming that the distributions are lognormal. This is a useful tool considering that number of samples reduces as

vehicle spacing increases which affects the variance of g. Since g is used as the link cost ε this would affect the prediction of η.

3.3.5 Optimal path

In order to improve the accuracy of the estimated η, the prediction of the first and last PV will rely on intermediary PV that are located in between the two vehicles. While an estimation can be made directly from the first to the last PV, it is expected that ε will be high. To minimize the ε, the problem is formulated as a shortest path (SP) problem.

SP is about finding the path between two nodes such that the sum of the weight of the links or ε in this case is minimized. When applied to the CF approach, SP is finding the path between a leader and the last follower such that the sum of ε is minimized. The leader and follower are the origin and destination nodes and the cost of a links is ε. The links connect the origin and destination nodes with other nodes, which are the intermediary PV. This can be visualized in Figure 16.

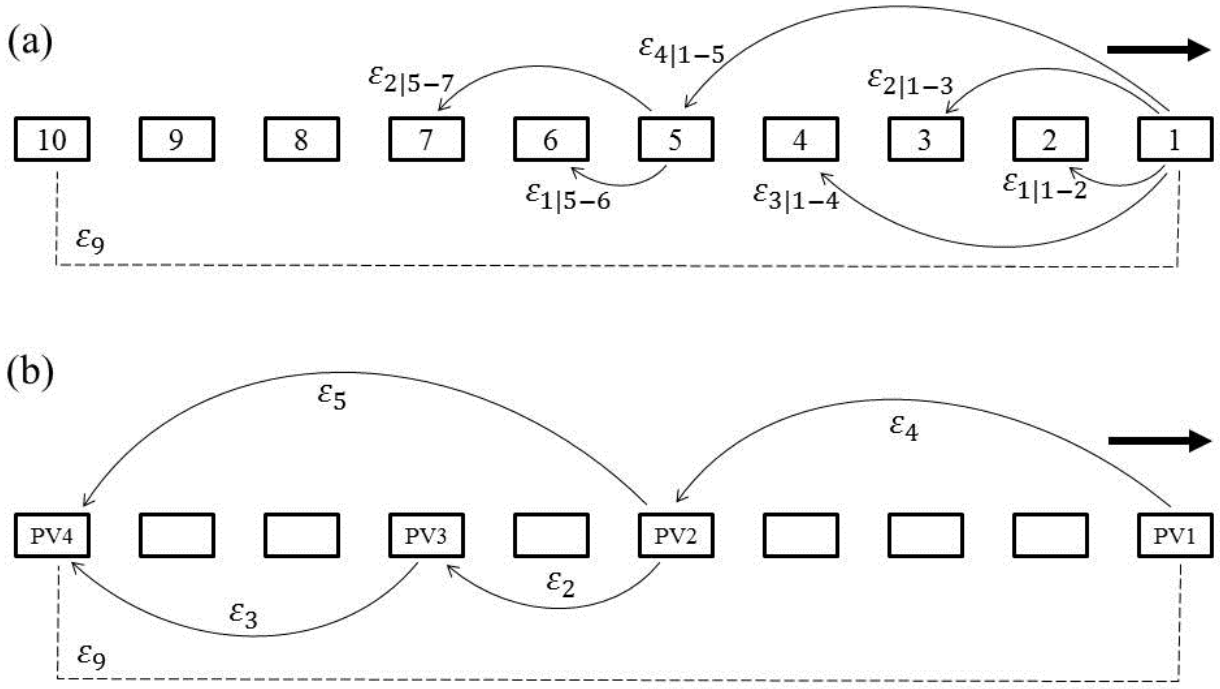

Figure 16 Probe vehicle shortest path problem

In Figure 16a, the square boxes are vehicles 1 through 10. From the training model, the ε for vehicle 1-2 is identified as $\varepsilon_{1|1-2}$ which is the error for η = 1. Followed by $\varepsilon_{2|1-3}$, $\varepsilon_{3|1-4}$, $\varepsilon_{4|1-5}$ continuing to $\varepsilon_{9|1-10}$ which are the errors for η = 2, 3, 4 and 9, respectively. It can also be observed that vehicle 2 is the leader for vehicles 3 through 10, vehicle 3 is the leader for vehicles 4 through 10, so on and so forth. To emphasize this point, in this figure vehicle 5 is the leader for vehicles 6 and 7 with $\varepsilon_{1|5-6}$ and $\varepsilon_{2|5-7}$, respectively.

Let's assume that there are four PV in the data as shown in Figure 16b. The objective is to predict η for PV4 with respect to PV1. During testing, the only data available are from these four PV. One option to estimate η is to calculate g between trajectories of PV1 and PV4. Then find

the highest probability of fitting a pdf from the training model, in this case with ε_9. Let's call this path A with ε_A.

Another option is to use the intermediary PV that are in between PV1 and PV4. It is expected that ε_2, ε_3, ε_4 and ε_5 to be smaller than ε_9. For this option, path B is PV1 – PV2 – PV4 with $\varepsilon_B = \varepsilon_4 + \varepsilon_5$ and path C is PV1 – PV2 – PV3 – PV4 with $\varepsilon_C = \varepsilon_4 + \varepsilon_2 + \varepsilon_3$. The shortest path problem will then pick the path with the smallest ε from a selection of ε_A, ε_B or ε_C.

To solve the SP problem, the well known Dijkstra's algorithm is implemented (Dijkstra, 1959). This algorithm is regularly implemented in route choice problems (e.g. transportation and computer networking). Today it is one of the most popular algorithms in computer science and is often applied in operations research and artificial intelligence. The steps in Dijkstra's algorithm are:

Step 1: Given a set of nodes and links, all nodes have infinite cost except for the origin, which has zero cost. Mark origin node as solved S.

Step 2: Determine candidate nodes that are connected to the origin node. From the weight of the link, calculate the cost between candidate nodes and solved node. Select the node with the minimum ε. Mark this node as solved and add to S.

Step 3: Update cost of solved node by weight of the link ε.

Step 4: Find unsolved nodes that are connected to V. The cost between unsolved nodes and S is the sum of the node cost and the link weight ε. Select node with the lowest cost. Update S.

Step 5: Update cost of solved node by weight of the link ε.

Step 6: Repeat step 4 and 5 until arrive at destination node.

To help understand Dijkstra's algorithm, the example shown in Figure 16b will be solved. First, the PV relationships are illustrated in terms of nodes and links. In this example, it is assumed that $\varepsilon_2 = 1$, $\varepsilon_3 = 2$, $\varepsilon_4 = 3$, $\varepsilon_5 = 4$ and $\varepsilon_9 = 8$. The initialization, node selections and final solution is illustrated in Figure 17.

In step 1 of Figure 17, the origin and destination nodes are identified. In this example, the origin is node 1 ending at node 4. In step 2, the candidates for node 1 are node 2 and 8 with ε of 3 and 8, respectively. The candidates are identified by the dashed lines. Node 2 is selected since it has the lowest ε. The cost for node 2 is updated and node 2 is added to S. In step 4, the candidate nodes are 3 and 4. The ε between node 4 – 2 is 4 and node 4 – 1 is 8. While the ε between 3 – 2 is 1. In step 5, node 3 is selected and it's node cost is updated. Step 6 is a repeat of the algorithm. The final solution of the shortest path problem is shown in step 7.

As seen from Figure 17, estimating η between the leader and the last follower can be viewed in terms of shortest path problem. By solving the shortest path problem, the model is able to predict η.

A heuristic approach called nearest node is introduced in finding the optimal path. This is an alternative to the SP method and is used as comparison. Typically, the node number in a graph network represent some arbitrary identification of a node. In this method, the node number represents the sequence of PV. The smaller the node difference, the nodes are closer to each other. In the nearest node approach, the optimal path is the path taken by utilizing the nearest node. Regardless of ε, the nearest node will always select the path to the nearest node until the target node is reached.

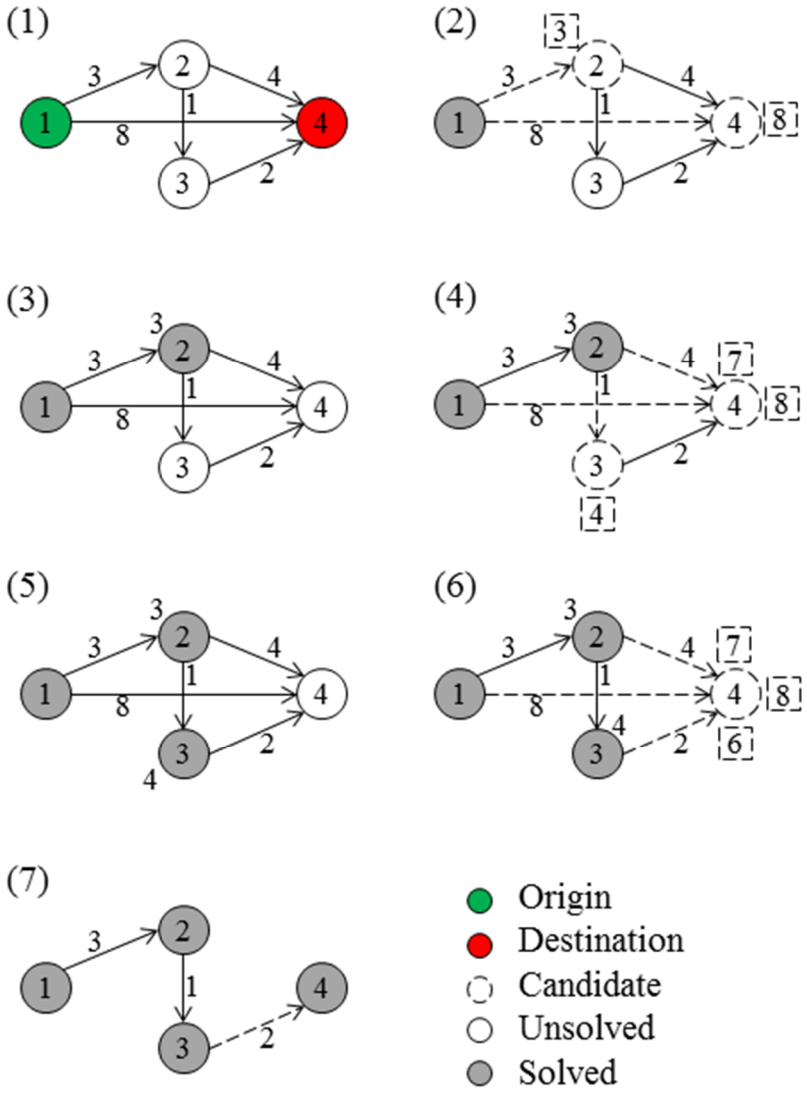

Figure 17 Solve shortest path using Dijkstra's algorithm

3.3.6 Link connection

In order to find the shortest path, there needs to be a complete link connecting the first and last PV. However, there are cases where the links are disjoint or intersect. This would make the prediction of η to be incomplete.

Disjoint is a case where the links are separated into groups. There are no links that connect the groups together. As a result, the predicted η using the SP would be for each group. To complete the process, η needs to be predicted for between the groups. Intersect is a condition where the links are overlapping each other. This results in duplication of η for the intersecting links. For a true prediction of η, the predicted η for the intersecting links needs to be removed. Figure 18 is an illustration of the complete, disjoint and intersect conditions.

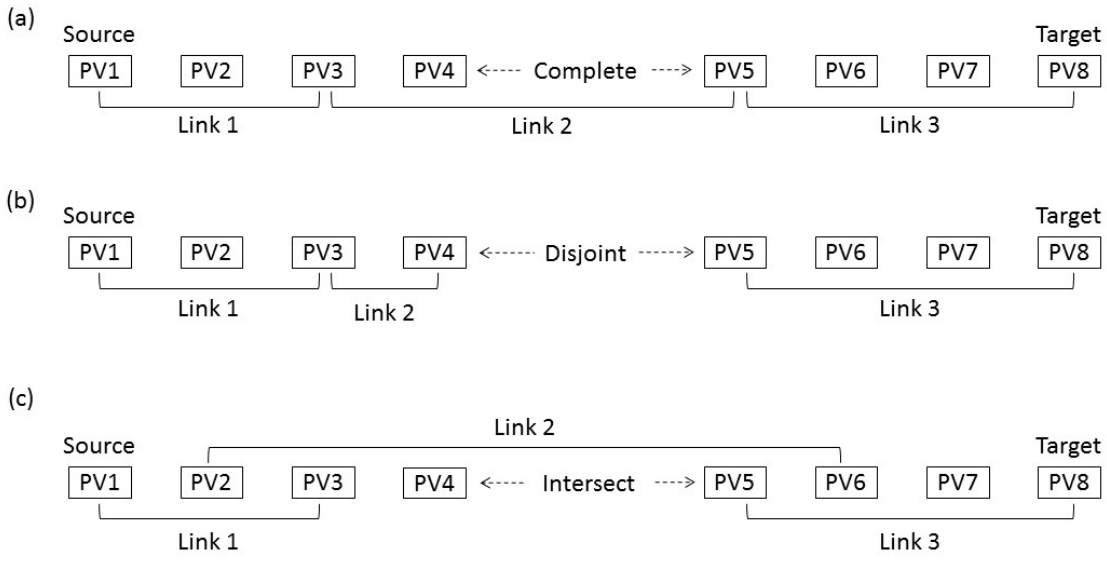

Figure 18 Probe vehicle link connection

Figure 18a illustrates a complete PV connection. In this figure, to go from PV1 (source) to PV8 (target), there exist a full connection. Link 1 connects PV1 and PV3, link 2 connects PV3 and PV 5 while link 3 connects PV3 and P8. With these three links, there is path from the source to target PV. Figure 18b is an illustration of a disjoint condition where link 1 connects PV1 and PV3 while link 2 connects PV3 and PV4. Link 3 connects PV5 and PV8. Even though there are links that connect the PV, there is no path that connects the source and target PV. The intersect condition is shown in Figure 18c. Link 1 connects PV1 and PV3, link 2 connects PV2 and PV6 and link 3 connects PV5 and PV8. In this figure, the links overlap each other without providing a path from the source to target PV.

To predict η for the disjoint condition, the last PV from the first group and the first PV from the subsequence group need to be identified. After identifying the PV, the time headway T between the two vehicles are calculated. T is then fitted to pdf of T for each η, which is developed from the training model. To estimate η, find the argument of the maxima from $f(T)$ which can be formulated as:

$$\hat{\eta} = \underset{\eta}{\operatorname{argmax}} f(T|\eta) \qquad 29$$

This method is similar to the method discussed earlier. The difference is that, instead of g, the variable T is utilized. The same for intersect condition, where the PV where the links overlap are identified. For disjoint, the η is added to the prediction for the groups. For intersect, η is deducted from the prediction of the groups.

From the training model, T is the difference in time where each vehicle in a pair is located at the same space. Obviously the vehicles cannot be in the same space at the same time or else there

would be a collision. Since data are collected every 1 second, the time increment would be the same for all vehicles, but not the space location. To find T, find the space where it is equal for both vehicles in a pair and calculate the time difference. In some instances, both vehicles do not have the same space and are therefore omitted from calculation. In other cases, there are more than one occurrences of both vehicles with the same space. When this occur the minimum of all T is designated as T. The pdf of T is then developed for each η.

From the test model, T is calculated by taking the last known time-space coordinate of the source PV. At this space, find the time for the target PV. Interpolation is performed if no direct value is available. T is the time difference between the source and target PV.

In summary, the framework of the CF approach can be represented as a flow chart, as shown in Figure 19.

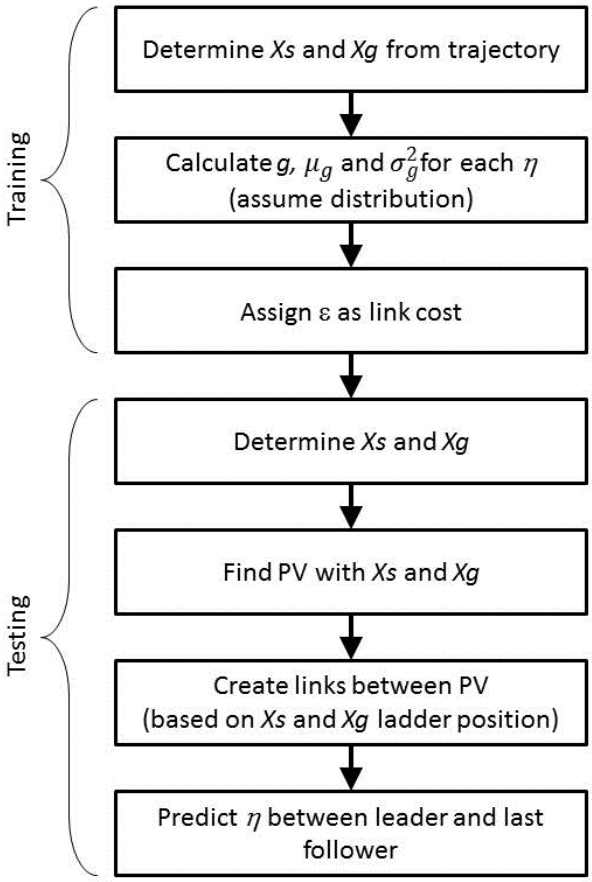

Figure 19 Framework for the car-following approach

3.4 PERFORMANCE MEASURES

Several different performance measures are used to evaluate the results of the proposed methodologies. The three indicators used are percent error (PE), mean absolute percentage error (MAPE) and root mean square error (RMSE). These three indicators measure the deviation of the estimated $\hat{q}$ or $\hat{\eta}$ and the observed q or η. They can be formulated as:

$$PE_i = \frac{F_i - O_i}{O_i} \times 100\% \quad (30)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{F_i - O_i}{O_i}\right| \times 100\% \quad (31)$$

$$RMSE = \frac{1}{n}\sqrt{n\sum_{i=1}^{n}(F_i - O_i)^2} \quad (32)$$

where: F_i is the i^{th} estimate value
O_i is the i^{th} observe value
n is the number of samples

These performance measures are selected because of their ability to capture deviations from different perspectives. PE measures the deviations between each observation which allows for further analysis of the data for identification of any trend, outliers, or similarities. MAPE and RMSE will each produce a single number that reflects the overall deviation of the data. These performance measures are commonly used and have been applied to other studies (Bucknell & Herrera, 2014; Kindzerske & Ni, 2007).

To evaluate the car-following approach, cross validation technique is implemented. Cross validation works by splitting the original data into a training set which is used to train the model and a test set to evaluate it. There are three different cross validation methods: random subsampling, *k*-fold and leave one out.

In random subsampling, the original data are randomly divided into groups: training and testing. Figure 20 is an illustration of this method.

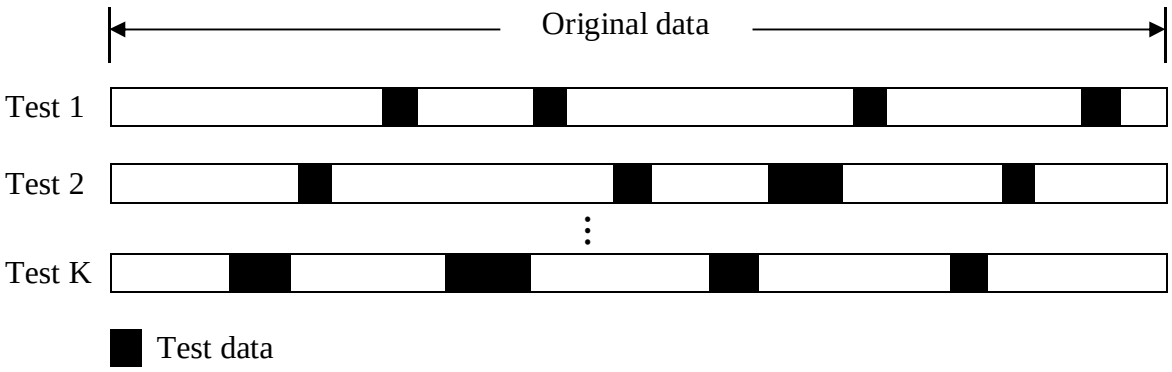

Figure 20 Random subsampling cross validation

In *k*-fold cross validation, the original data are divided into *k* groups. Of the *k* groups, a single group is used as test data while the remaining *k* -1 groups are used as training data. The cross validation is then repeated *k* time (or folds) with a single group used as a test data only once during the process. Figure 21 is an illustration of this method.

The final cross validation method is called leave one out (LOO). This method is the extreme approach to the *k*-fold method. Instead of folding for every group, LOO would fold for every observation from the original data, as shown in Figure 22.

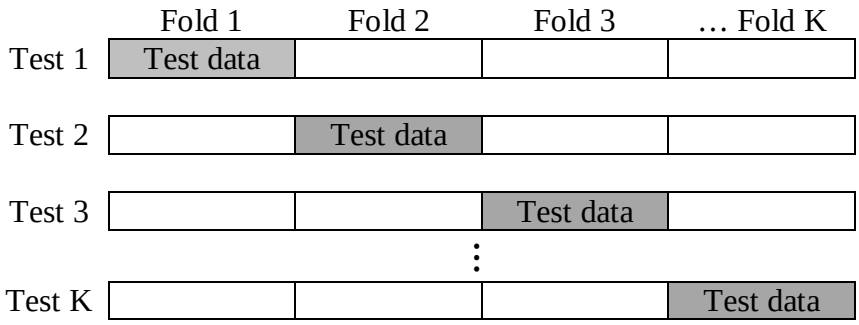

Figure 21 *k*-fold cross validation

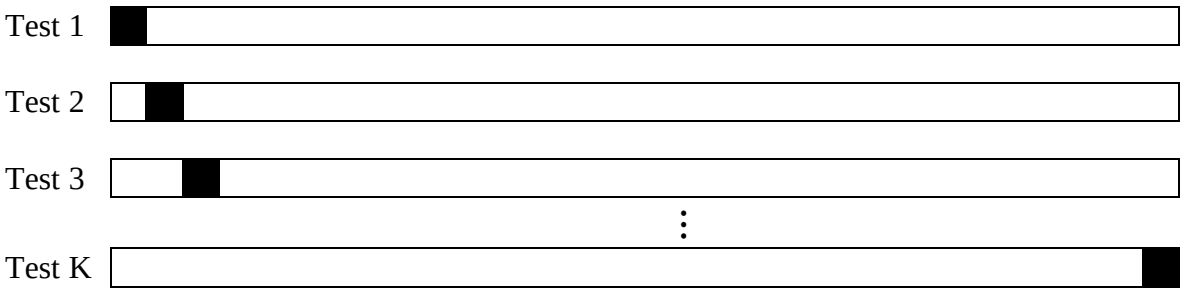

Figure 22 Leave one out cross validation

For the car-following approach, the *k*-fold cross validation method will be utilized. The implementation of the *k*-fold cross validation will be discussed under the Results chapter. This concludes the Methodology chapter. The following chapter is a discussion on the data used for analyses.

CHAPTER 4

4 DATA

Different sets of data are either acquired or generated for this dissertation. For the FD approach, data from a case study named Mobile Century is used. The shockwave approach utilizes simulated data of a hypothetical network while the car-following approach is developed based on the well known NGSIM data. This chapter explains all three types of data sets.

4.1 MOBILE CENTURY

For the FD approach, data from field observations are implemented. The study site is a corridor along I-880 located near Union City between Stevenson Boulevard and Winton Avenue ramps in the San Francisco bay area. Data were collected on Friday February 8th 2008 between 10:00am to 6:00pm. This data collection effort was a part of the Mobile Century project (Juan C. Herrera et al., 2010).

In this experiment, 165 drivers were recruited to travel specific routes along the study site. The drivers were equipped with GPS-enabled cell phones which, on average, transmitted data every three seconds. Data collected were latitude, longitude and timestamp. Post processing by the Mobile Century team added speed and postmile to the PV data. The penetration rate of PV is in the range of 2-5% depending on time of the day.

The Mobile Century project covered a corridor of approximately eleven miles in length in each southbound and northbound direction. The corridor is a four lane facility in each direction occasionally expanding to five lanes for merging lane near ramps. Upon closer inspection of the

data, it was noticed that at specific postmiles the penetration rate of PV was considerably low in the afternoon hours. The data from this time space region are excluded from analyses.

The Mobile Century team reported a mid-morning crash in the northbound direction between postmile 26 and 27. The team mentioned that recurring congestion occurs in the northbound direction. Considering the factors above a smaller segment around postmile 25 in the northbound direction was selected for this study. Figure 23 shows the entire study site and the PV trajectory between postmile 21 and 27 with an inset of the PV trajectory during the mid-morning crash.

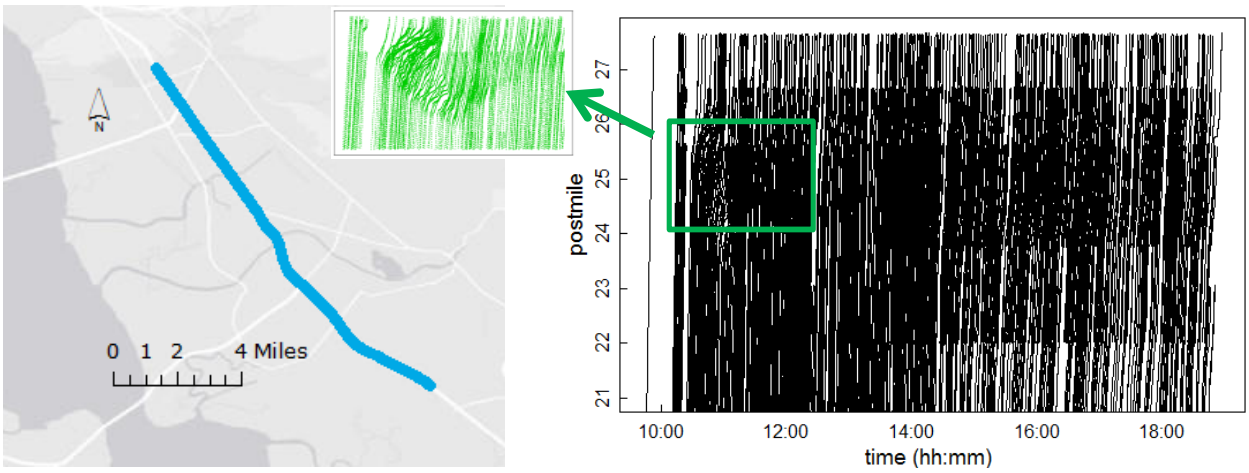

Figure 23 Study site (left), vehicle trajectory (right) and trajectory during incident (insert)

In addition to PV data, the experiment also collected loop detector data along the study site. The data collected were q and occupancy for each lane aggregated every 30 seconds for a duration of eight hours. As a supplement, additional data were retrieved from the California Department of Transportation PeMS website for the same study area. The downloaded data included q and u for each lane aggregated every five minutes for 34 days starting from January

6th to February 9th, 2008. For eight hours on February 8, 2008 there is an overlap of loop data from Mobile Century (in 30 second increments) and PeMS (in 5 minute increments). Analysis shows that there was a match of traffic flow between both datasets during this eight hour time period. Due to limited amount of loop detector data from the Mobile Century, only PeMS data are used for this approach.

To calibrate the FD to the PeMS data, k must be known. For this approach k was calculated using Equation 11. With the exception of Van Aerde, the fitting of the FD to the PeMS data was performed by regression using statistical software R. To determine the parameters for the Van Aerde FD, a software was developed to perform such calculation (Michel Van Aerde & Rakha, 1995). A user inputs the values of q, u and k and the software will calculate u_c, u_f, q_c and k_j. The regression formulas are shown in Table 1.

After calculating all of the parameters, k can be calculated for any u. In this dissertation, u started from zero and was incrementally increased by one until a maximum value of ninety. k was then calculated for every u. Due to the different formulation of $u-k$ relationship for any given u, the value of k will vary for each FD, resulting in a unique $u-k$ shape for each FD. q is then calculated by multiplying u and k. By knowing all three values, the $u-q-k$ relationship for each FD can be established.

Figure 24 plots the observed q, u and the calculated k from PeMS data along with the fitted FDs. In this figures the grey points represent the observed values. The four different FDs are plotted using different colors and line types. The R^2 values for the fitted FDs from the $u-k$ relationship are 87%, 85% and 94% for Greenshields, Underwood and Northwestern,

respectively. The Van Aerde FD was fitted to the $u - q$ relationship and no R^2 value was available.

Table 1 Fundamental diagram relationships

Model	Speed-Density Relationship	Regression
Greenshield	$k = k_j\left(1 - \frac{u}{u_f}\right)$	$k = k_j - \frac{k_j}{u_f}u$
Underwood	$k = k_o \ln\left(\frac{u_f}{u}\right)$	$k = k_o \ln u_f - k_o \ln u$
Northwestern	$k = k_o\left(2\ln\frac{u_f}{u}\right)^{1/2}$	$k^2 = 2k_o^2 \ln u_f - 2k_o^2 \ln u$
Van Aerde	$k = \frac{1}{c_1 + \frac{c_2}{u_f - u} + c_3 u}$	$n = \frac{2u_c - u_f}{(u_f - u_c)^2}$, $c_2 = \frac{1}{k_j\left(n + 1/u_f\right)}$, $c_1 = nc_2$, $c_3 = \frac{-c_1 + \frac{u_c}{q_c} - \frac{c_2}{u_f - u_c}}{u_c}$

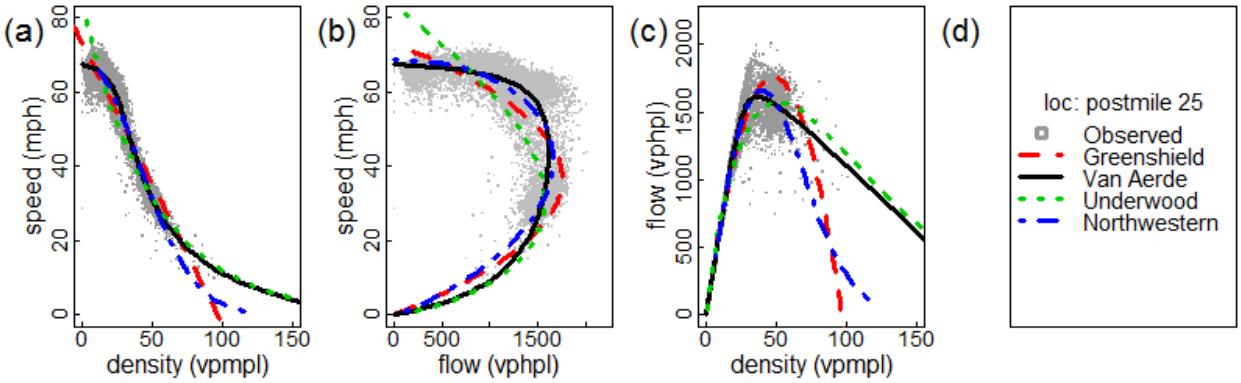

Figure 24 Different models of fundamental diagrams

In Figure 24a and c, the FDs, with the exception of Greenshield, show high values of k as u and q approaches zero. This is due to lack of observed points at low speed which affects the regression results. In Figure 24b, Underwood shows an unrealistic u_f consistent with findings from May (1990). Even though the focus of this dissertation is on the $q - u$ relationship, it is important to understand the $u - k$ regression results because q is derived after u and k. Overall the FDs fit the observed values with different FDs taking on different shapes.

To compare $\hat{q}$ from the FD to observed q, data standardization between PV and PeMS speed is required. Since loop detector location is fixed while PV move in space, the first step is spatial standardization. To standardize the data a loop detector must first be selected. After identifying a loop detector, a 0.05 mile area upstream and another 0.05 mile area downstream of the loop are designated as a target area. Any PV data within this area are used for analyses. A 0.1 mile target area was selected because this would guarantee at least two PV points within this area, which is the minimum number of observations required for further analyses, e.g., regression. This was under the assumption of PV free-flow speed of 75 mph and data collection of every three second.

Expanding the area even more could infringe upon upstream or downstream traffic flow characteristics, e.g., on/off ramps.

Temporal standardization is the second step of the process. PeMS data are in 5-minute aggregation while PV data are updated every three to four seconds. To standardize the PV data, all of the PV speed in the target area are aggregated according to the PeMS time interval. Figure 25 illustrates a sample of spatial and temporal standardization for a thirty-minute period from 10:30am to 11:00am.

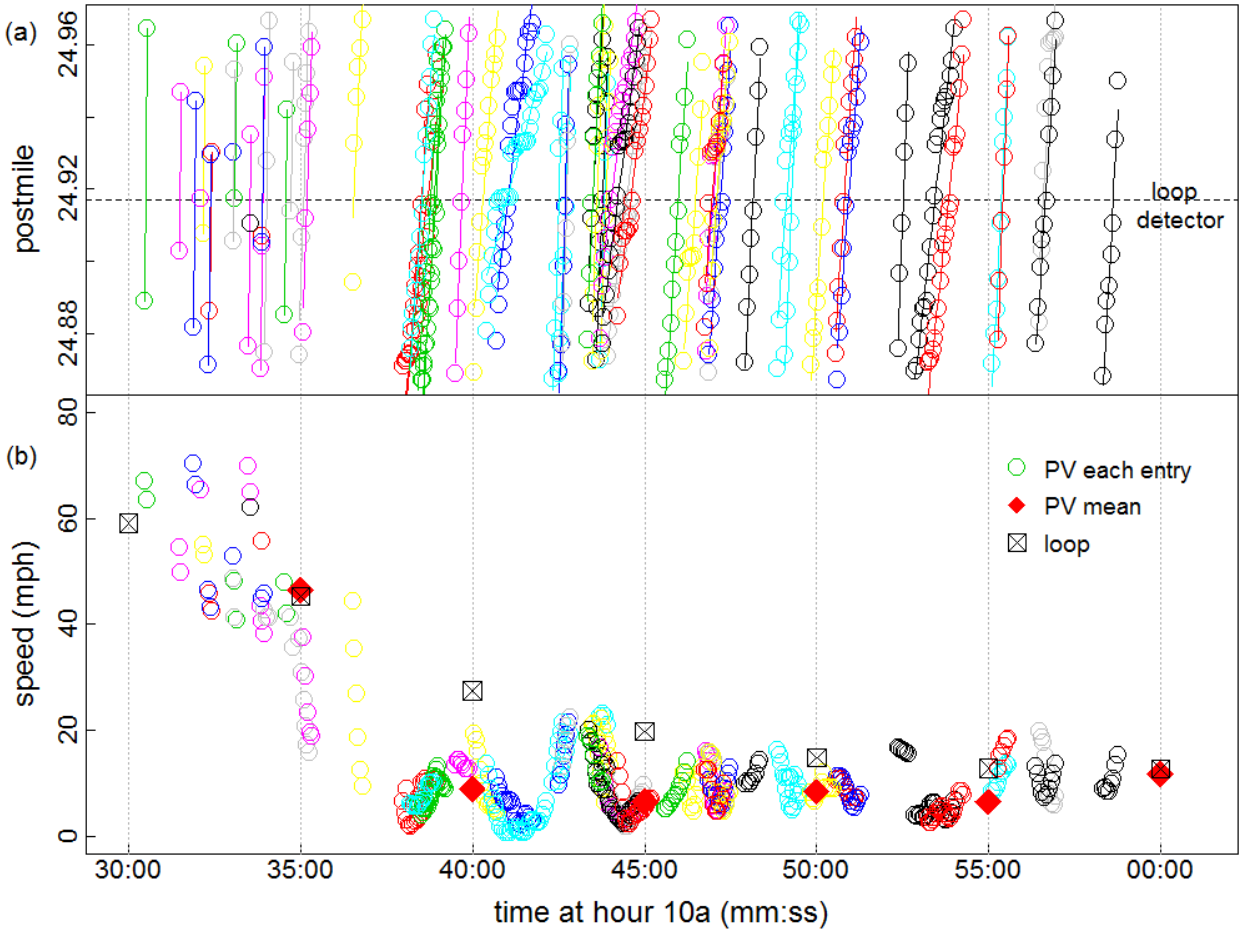

Figure 25 Probe vehicle speed (5-min aggregation)

Figure 25a shows the spatial standardization where the horizontal dotted line is the loop detector location and the colored circles are PV readings. Different color circles indicate different PV. The solid lines connecting the circles are linear regression lines. At the beginning of 10:30am, most of the PV have on average two readings as they travel through this area. This indicates that PV are travelling at free-flow speed. As time progress more points for each PV can be observed in this area, which indicates slower speed.

Figure 25b plots the PV speed against time for the target area. Colored circles indicate the PV speed for each observation with different color circles signifying different PV. The square x's are the aggregated speed recorded from the loop detector at 5-minute intervals. The diamonds are the aggregated 5-minute PV speed. From this figure the PV and loop speed are almost the same at the beginning (free-flow) and at the end (congestion). As traffic transitions from free-flow to congestion around 10:40am, there is almost a 20 mph difference between PV and loop detector speed. Several possible reasons for the difference in speed are: (1) lane to lane traffic flow variation with majority of the PV travelling at the slow moving lane, and (2) PV represent a small sample of the flow while the loop is collecting data for all vehicles. As traffic flow becomes uniform, in this case towards congestion, the difference between PV and loop detector speed is reduced.

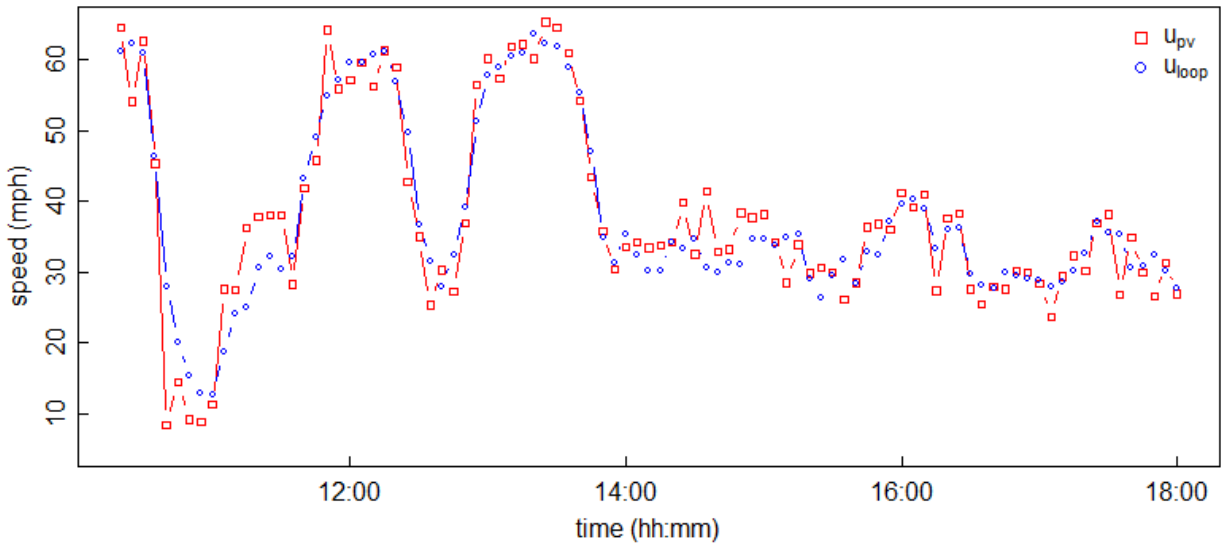

Figure 26 Probe vehicle and loop detector speed comparison

Due to the limited time period of thirty minutes, a longer time period is beneficial to understand the relationship between PV and loop detector speed. Figure 26 compares the PV and loop speed in 5-minute aggregation for the entire eight-hour period. While there are differences between the two readings, overall, both PV and loop detector speed follow the same trend.

Detail analyses of the Mobile Century data suggest that, overall, the PV and loop detector data can be utilized for estimating $\hat{q}$. In the following section, data generated from simulation will be discussed.

4.2 SIMULATION

Simulation data are applied to the shockwave approach. To generate data, a road segment is simulated in Vissim as shown in Figure 27. The segment is a three lane freeway facility reducing to a two lane segment at the end of the roadway. The segment length is over three miles. No

ramps exist between the start and end points of the segment. This segment is not intended to simulate any actual road segment. It is a hypothetical segment with the aim of producing synthetic data to be applied to the shockwave approach.

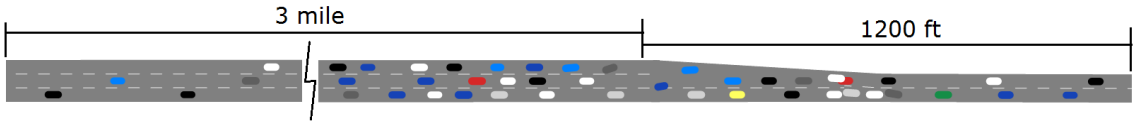

Figure 27 Vissim simulated network

To create congestion a bottleneck from three to two lanes is intentionally introduce with a high demand in the beginning and slowly decreasing towards the end. Table 2 lists the simulated demand. From visual observation the queue from the bottleneck never reaches the starting point of the segment. Simulation time was two hours with a total of 6,669 vehicles. All other simulation parameters (e.g., car following, free-flow speed) are based on Vissim default values.

Table 2 Simulated demand

Simulation time (sec)	0-900	900-1800	1800-2700	2700-3600	3600-4500	4500-5400	5400-6300	6300-7200
Demand (vehicle)	4000	4600	4400	4200	3500	3000	3000	2500

Figure 28 is a randomly selected trajectory of five percent of the overall vehicles. Each line is the trajectory for a single vehicle. Lines that are near vertical indicate free-flowing condition.

Angled lines indicate congestion. From visual observation a backward moving shockwave can be observe between time zero to about 4000 second. As demand decreases, the shockwave changes to forward moving starting around 4000 second and finally disappearing around 6500 second.

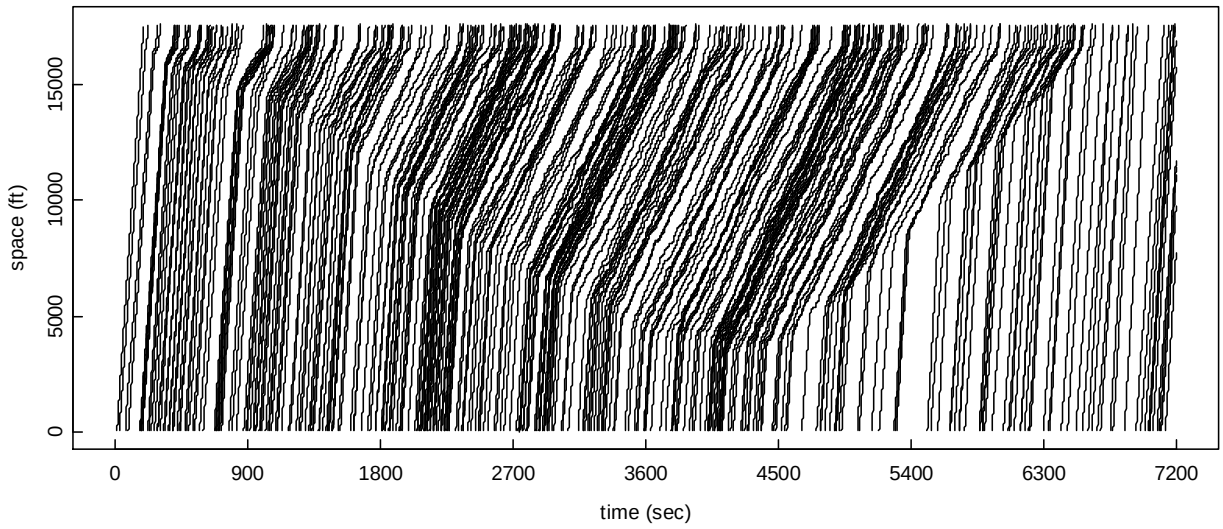

Figure 28 Simulated probe vehicle trajectory

From the simulated data of all vehicles, congested flow q_j and jam density k_j can be determined. To calculate q_j the number of vehicles is counted at location 16000 feet between 1800 and 5400 second time period (equivalent to 1 hour). k_j is estimated by counting the number of vehicles at time 3600 second between 10000 and 15280 feet (equivalent to 1 mile). From calculation of all vehicle at the specified time-space, q_j and k_j were determined to be 1,159 veh per hour per lane (vphpl) and 128 veh per mile per lane (vpmpl), respectively. These values are the ground truth and will be used in other calculations.

To classify the free-flow and congested periods *k*-means clustering is applied to the simulated data. Three different variables, which are PV time, location and speed, can be used to

perform the cluster analysis. Time and location are independent variables while speed is the distance traveled over a period of time. Since speed is dependent upon location and time, it is chosen as the variable used for clustering. For this analysis the number of clusters *k* is set to two. Assuming a five percent PV penetration rate, Figure 29 illustrates the results from *k*-means clustering along with free-flow speed u_f and congested speed u_j. The free-flow period is colored green while congested period is blue.

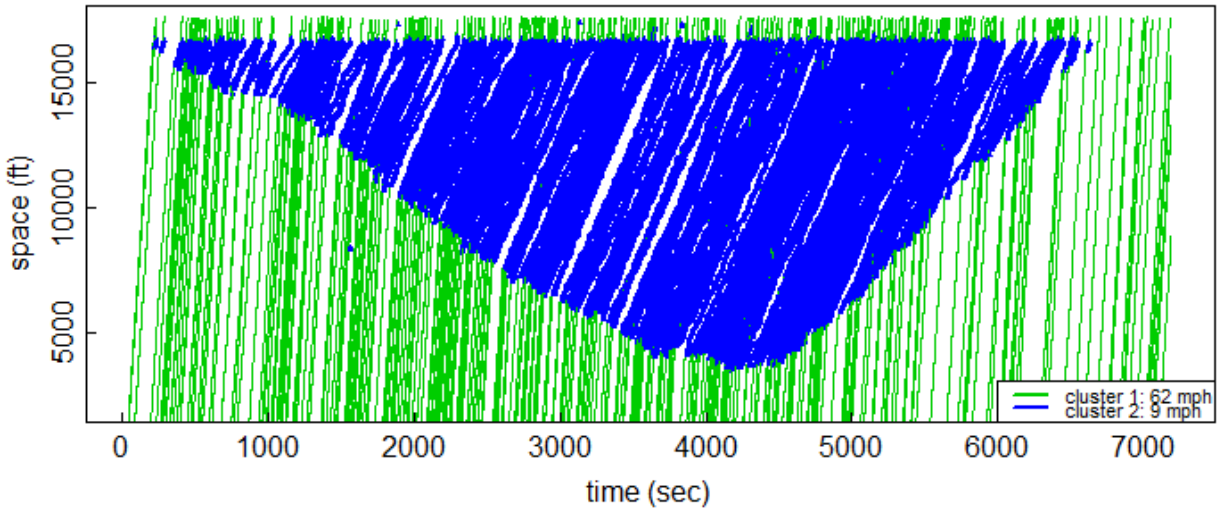

Figure 29 Results of *k*-means clustering

In general, *k*-means clustering is able to classify the free-flow and congested regions. It is noticeable that small clusters of congested regions are observed within free-flow region. From Figure 29 shockwave can be observed between free-flow and congested regions. Closer inspection reveals that the shockwave is not a constant value throughout the simulation but

instead changes over time. This is due to different demand or arrival rates during different times of the simulation.

As stated earlier in this section, demand is increased at the beginning of the simulation before decreasing towards the end. In the simulation the demand is modified every 900 seconds or 15 minutes. Throughout the simulation the demand is modified eight times that can be identified as time segment i. To calculate ω_i for each time segment, the shockwave equation is used where:

$$\omega_i = {(q_j - q_i)}/{(k_j - k_i)} \quad (33)$$

where: ω_i is shockwave for time segment i
q_i is flow for time segment i
k_i is density for time segment i
q_j is flow during congestion (calculated to be 1159 vphpl)
k_j is jam density (calculated to be 128 vpmpl)

q_i is the number of all vehicles entering the network at each time segment i. k_i is the density of vehicle each time segment. k_i is calculated at middle of each time segment for a one half mile during free-flow period. Multiply by two to convert k_i to vehicle per mile. After calculating q_i and k_i combined with the calculated q_j and k_j, ω_i can now be calculated. The resulting ω_i and q_i are shown in Table 3.

Table 3 Simulated flow and shockwave

Time Segment	1	2	3	4	5	6	7	8
Time (sec)	0	900	1800	2700	3600	4500	5400	6300
Demand (vph)	4000	4600	4400	4200	3500	3000	3000	2500
q_i (vph)	4071	4401	4323	4179	3477	2133	1980	2103
ω_i (mph)	-1.90	-2.80	-2.71	-2.17	0	3.80	4.38	3.95

From Table 3, ω_i is backward moving (negative) in time segments 1 through 4. In time segment 5, ω_i is transitioning from backward to forward resulting in a zero value. As the simulation concludes ω_i is forward moving (positive). ω_i shown in Table 3 will be used as ground truth to determine u_B, which is the breakpoint speed as proposed by Northwestern congested regime FD.

4.3 NEXT GENERATION SIMULATION (NGSIM)

NGSIM is a data collection effort championed by the Federal Highway Administration (FHWA) of the US Department of Transportation. The data consist of vehicle trajectory, detector and other supporting data (e.g. weather, road sign) for several different roadways. Vehicle trajectories are recorded from video cameras which are mounted on top of tall structures along the roadways. From the video images, vehicle positions are identified for every tenth of a second.

There have been countless number of studies based on NGSIM data. Lu and Skabardonis (2007) developed a four-step algorithm to predict *w* based on vehicle trajectory. From the NGSIM I-80 and US 101 datasets, they've determined *w* to be 11.4 mph (16.8 foot per second). Izadpanah, Hellinga, and Fu (2009) proposed an alternative algorithm to predict *w*. When

applied to US 101 dataset *w* was calculated to be 13.8 foot per second. That's a difference of 17.5% compared to the estimated value from Lu and Skabardonis (2007).

Treiber, Kesting, and Thiemann (2008) applied the NGSIM vehicle trajectory data to the fuel consumption model that they've developed. The model can then be implemented to a simulation software to calculate fuel consumption and emission. Just like any data, the NGSIM vehicle trajectory data is not immune to error. Punzo, Borzacchiello, and Ciuffo (2011) proposed a method to assess the accuracy of vehicle trajectory involving jerk, consistency and spectral analysis.

The NGSIM data collection took place at several locations accounting for different types of facility (e.g. freeway, arterial) and characteristics (e.g. merge/diverge). In this dissertation, data collected from I-80 and US 101 are used. The I-80 study site is located in Emeryville, California for northbound traffic. The study area is approximately 500 meters (1,650 feet) in length. It consists of five mainline lanes with an additional 6^{th} lane dedicated to the Powell Street and Ashby Avenue on/off ramps. The left most lane is the high occupancy vehicle HOV lane. Data were collected on April 13 2005, divided into three 15-minute periods: 4:00 pm to 4:15 pm, 5:00 pm to 5:15 pm and 5:15 pm to 5:30 pm. Seven synchronized video cameras were used to record the vehicles. Figure 30 illustrates the I-80 study area.

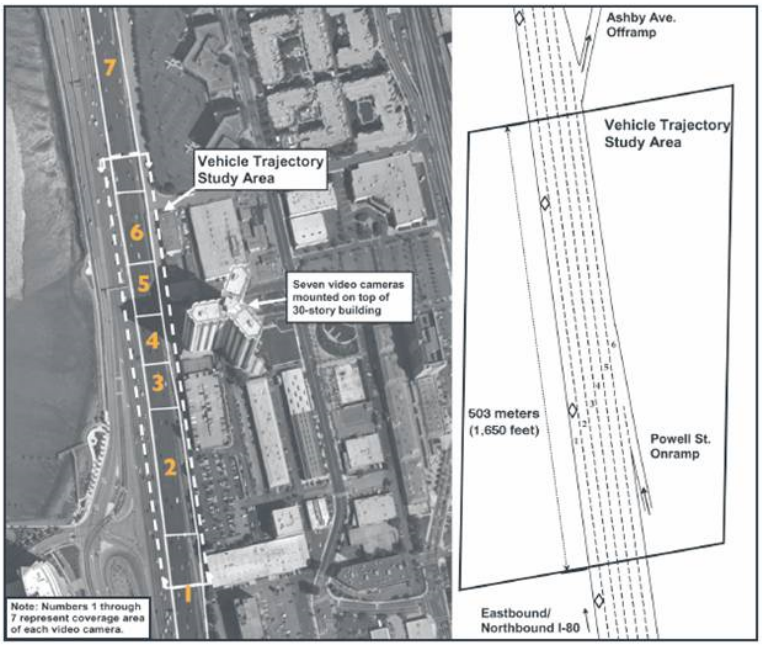

Figure 30 NGSIM I-80 study site

The other study site is the US 101 located in Los Angeles, CA for southbound traffic. Data were collected on June 15 2005, divided into three 15-minute periods: 7:50 am to 8:05 am, 8:05 am to 8:20 am and 8:20 am to 8:35 am. This study site is 640 meters in length consisting of five mainline lanes and an additional 6th lane serving the Ventura Boulevard and Cahuenga Boulevard on/off ramps. Eight synchronized video cameras were used to record the vehicles. Figure 31 is an illustration of the US 101 study area.

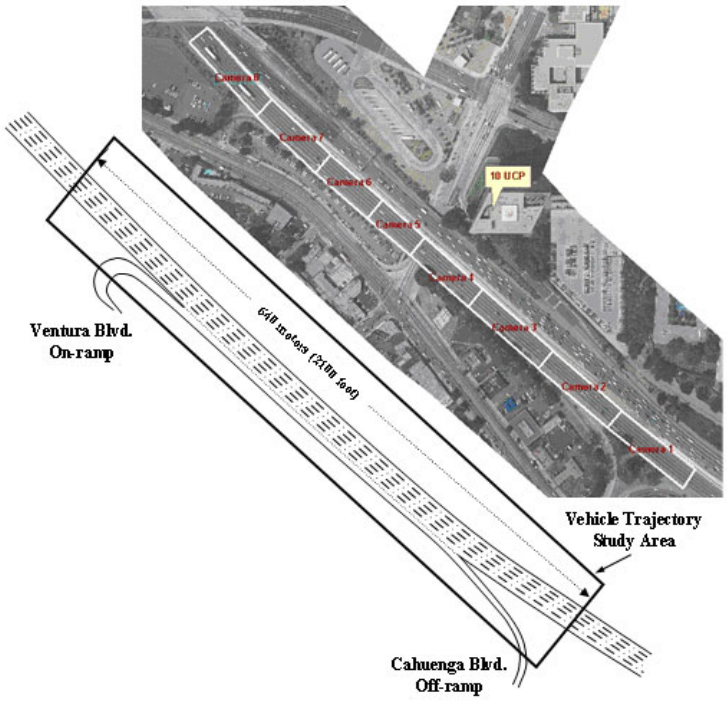

Figure 31 NGSIM US101 study site

Format of data for both study sites are the same. The vehicle trajectory data had 18 variables, listed as:

1. Vehicle identification number
2. Frame identification number
3. Total number of frames in which the vehicle appears in the data set
4. Elapsed time since January 1 1970 (millisecond)
5. Lateral coordinate of the front center of the vehicle with respect to the left-most edge of the section in the direction of travel (feet)
6. Longitudinal coordinate of the front center of the vehicle with respect to the entry edge of the section in the direction of travel (feet)

7. X Coordinate of the front center of the vehicle based on CA State Plane III in NAD83 (feet)
8. Y Coordinate of the front center of the vehicle based on CA State Plane III in NAD83 (feet)
9. Vehicle length (feet)
10. Vehicle width (feet)
11. Vehicle class (1-motorcycle, 2-auto, 3-truck)
12. Vehicle speed (feet per second)
13. Vehicle acceleration (feet per second square)
14. Lane position of vehicle
15. Vehicle identification of the lead vehicle
16. Vehicle identification of the follower vehicle
17. Space headway
18. Time headway

The strength of NGSIM data is the completeness of data collection from lane assignment of the vehicles, to the vehicle's follower and leader recorded at a high frequency. Such complete data allows for a detail analysis of the vehicle trajectory. On the other hand, the weaknesses of this dataset are the limited time period (45 minutes) and relatively short segment lengths. Even with the weaknesses, NGSIM is still an attractive dataset for analyses.

As shown from the aerial views, there are on and off ramps within the study areas. With the presence of the ramps, significant lane changes occurred. This is bound to happen as vehicles

exit or enter the mainlines. Which would then affect the vehicles on the mainlines as they change lanes in reaction to the exiting/entering vehicles from the ramps. To show the lane changing behavior, Figure 32 is the lane assignments of the vehicles as they travel through I-80 (a) and US 101 (b). The data for I-80 are from 5:00 pm to 5:30 pm and data for US 101 are from 7:50 am to 8:35 am.

In Figure 32, the lines are trajectories of the vehicles going through the study areas. The purpose of these figures is to show the lane changing behavior of the vehicles, not to identify which vehicles are making the lane change and which vehicles are not. In these figures lines that are crisscrossing between the lanes indicate lane changing. It can be observed that lane changing is heaviest near the ramps. Less lane changing occurs further away from the ramps. However, if the starting points of the study area were to move back upstream, it is possible that lane changing is heaviest further away from the ramps. This could occur as vehicles prepare to move from the mainlines and head towards the exit lane.

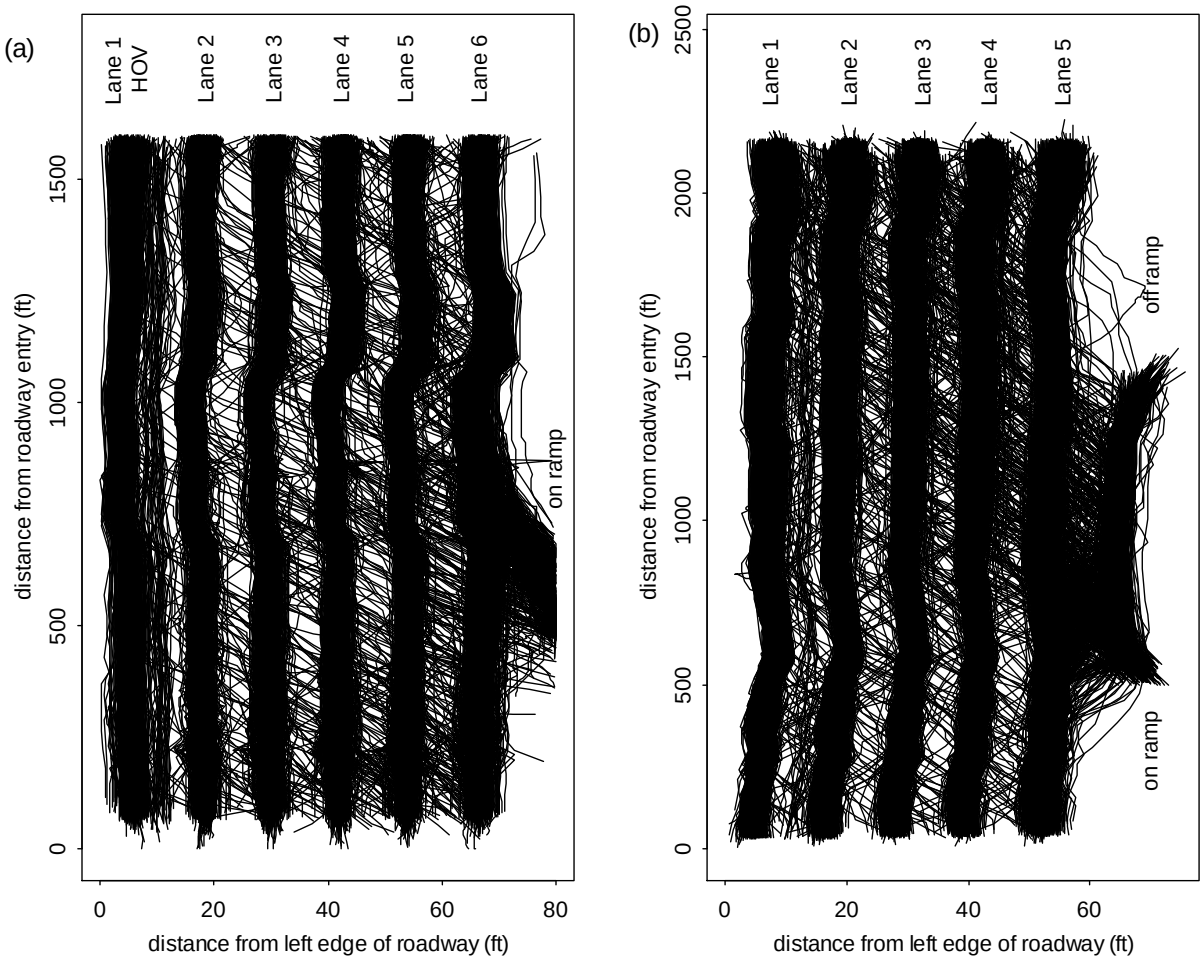

Figure 32 NGSIM vehicle lane assignment (a) I-80 and (b) US101

In addition to lane changing behavior, Figure 32 also shows some anomaly with the data. For example, there are several vehicle trajectories located to the right of lane 6 on I-80. These trajectories are located on the upper right hand corner of Figure 32a. The anomaly of these vehicle trajectories could be due to error in data processing or vehicles traveling on emergency lane.

To understand the dynamics of speed of traffic, a speed heatmap for each lane for both sites are developed. Figure 33 and Figure 34 are the speed heatmaps for I-80 and US 101,

respectively. The I-80 heatmap is from 5:00 pm to 5:30pm while the US 101 heatmap is from 7:50 am to 8:30 am. In the heatmaps, black color indicates a speed of zero and green indicates a speed of 70 mph. The color palette changes from red to orange to yellow as speed changes from 0 to 70 mph.

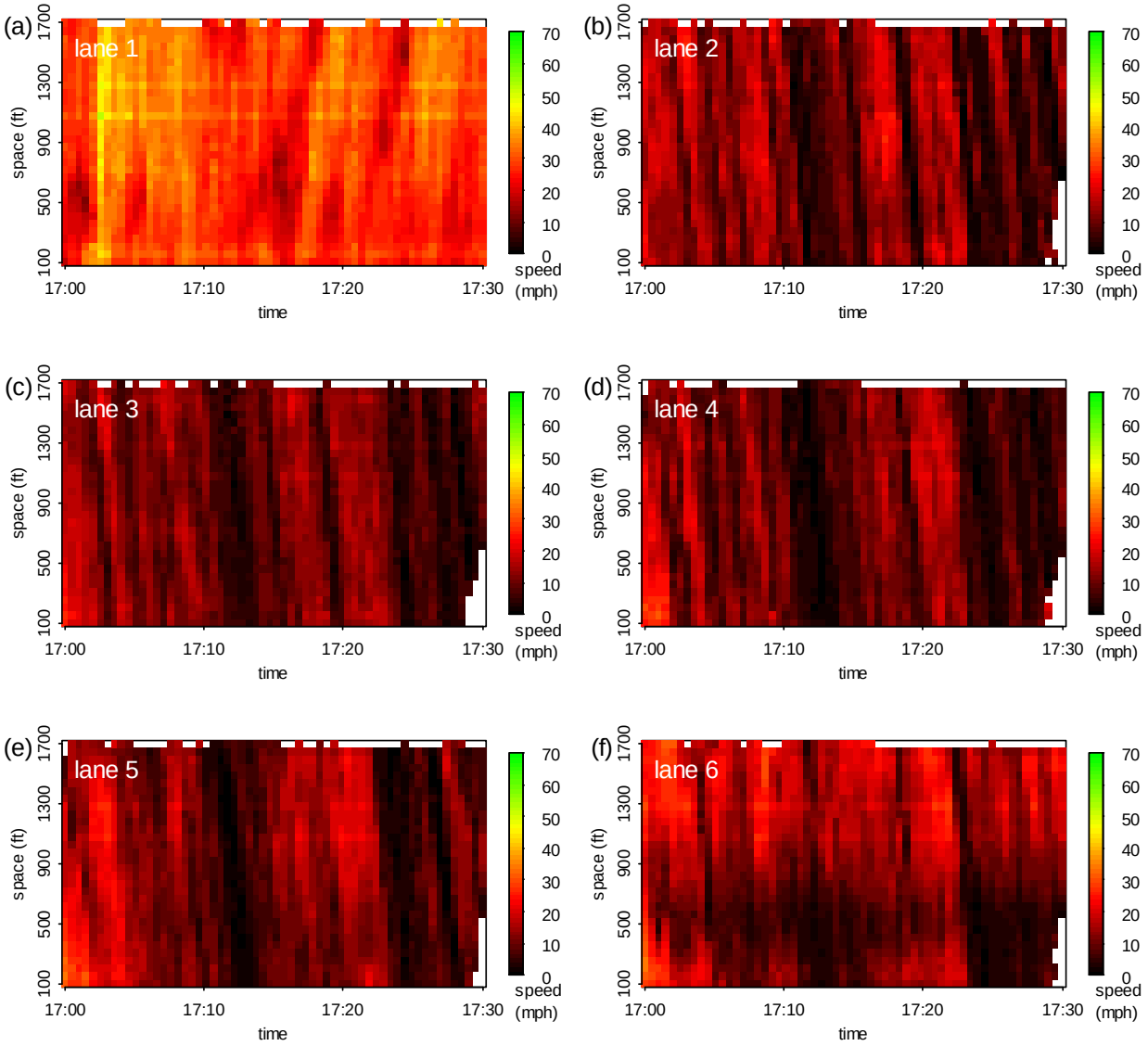

Figure 33 I-80 speed heatmap

For I-80, lane 1 which is the HOV lane exhibits the least amount of speed fluctuation. This is expected due to the restrictions put in place for this lane. As the name indicates, HOV lane requires that a vehicle must have more than one person travelling in it. To make car pooling attractive, it is imperative that HOV lane has an advantage over non-HOV lanes. From the heatmap, it can be observed that the speed for lane 1 is higher than the other lanes.

As for the remaining lanes 2 through 6, there is considerable amount of stop and go traffic as the heatmap color alternates between black and red, never reaching yellow or green. While the pattern between each lane is different, overall lanes 2 through 6 show similar speed pattern.

There is no HOV lane for US 101. As a result, the speed patterns for all lanes on US 101 are almost the same, as shown in Figure 34. While there are stop and go traffic as indicated by the black color, majority of the traffic speed is in the 30 mph range as indicated by the reddish orange color. Interestingly enough, the black color on the heatmaps could also indicate the shockwave of the stop and go traffic.

When comparing the speed heatmaps between I-80 non-HOV lanes and US 101 lanes, it is observed that congestion on I-80 is more severe than US 101. This is based on the observation that the speed heatmaps for I-80 are primarily reddish black compared to reddish orange for US 101.

Due to vehicle merging and diverging, a smaller set of data are extracted from the original NGSIM data. The lane changing of the vehicles presents a challenge for analyzing the data. To mitigate this issue, a smaller set of data are extracted at which no lane change occurred. To extract this data, an area near the end of the segment is selected. In this area, the lane that is connected to the on-ramp at the start of the segment is leading into an off-ramp. It is expected

that vehicles are in their desired travel lanes within this area. More vehicles in this area would remain in their lanes because the merge/diverge had already occurred upstream from this location.

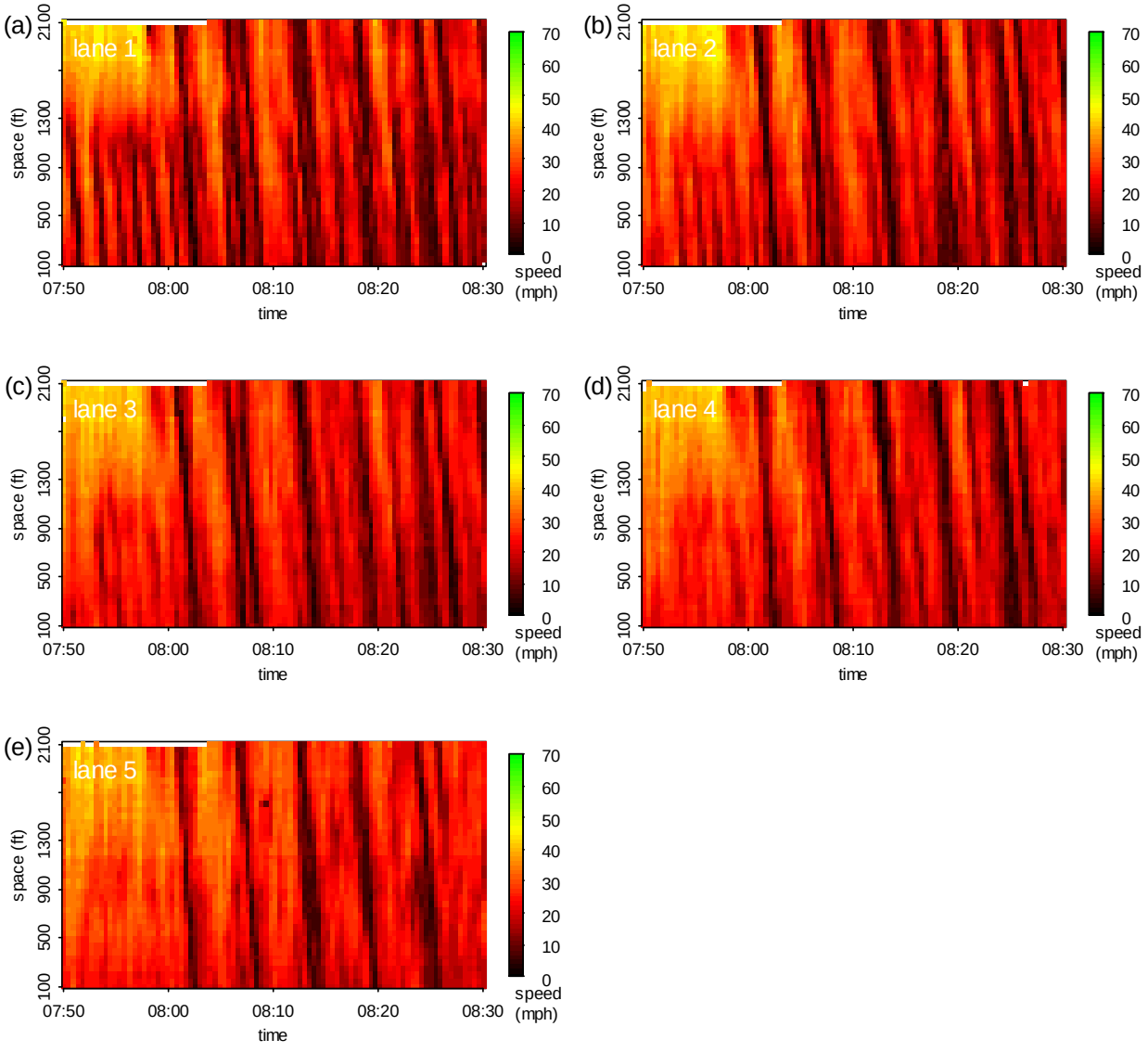

Figure 34 US101 speed heatmaps

After narrowing down the search area, data are scoured during different time periods. Lane changing vehicles are excluded from the search. As a result, there is a small window where no lane change occurred. For US 101, that window is from 900 to 2100 feet and approximately between 8:14 am and 8:16am. For I-80, the window is from 1200 to 1600 feet and approximately between 5:10 pm and 5:13 pm. The vehicle trajectory for the search area are shown in Figure 35. Had any of the vehicles change lane, the trajectory of the vehicle would cross each other. However, since no trajectory crosses one another, it can be concluded that vehicles did not change lane in this time-space region. Another way of confirming the lane change is by analyzing the variable "lane position of vehicle" which is a part of the data collection. To check for lane changes, simply ensure that the values for lane position do not change.

In addition to lane changing movements, the data are also filtered for the vehicle class. For analyses, only vehicles classified as passenger cars are considered. Other vehicles such as trucks or motorcycles would make analyses complicated and are therefore omitted from analyses. The trajectory shown in Figure 35 are for passenger cars only.

Figure 35a through c are the vehicle trajectory for US101 and figures d through f are for I-80. In Figure 35c, the vehicle trajectories are truncated due to lane changing towards the end of the study period. If these lane changing vehicles are removed completely from the data, it would severely reduce the number of vehicles or sample size. Therefore, the decision is to truncate the data rather than complete removal. Lane 4 and 5 of US101 and lane 3 and 6 of I-80 are omitted from analyses due to lane changing or presence of trucks. Lane 1 of I-80 is an HOV lane and also omitted.

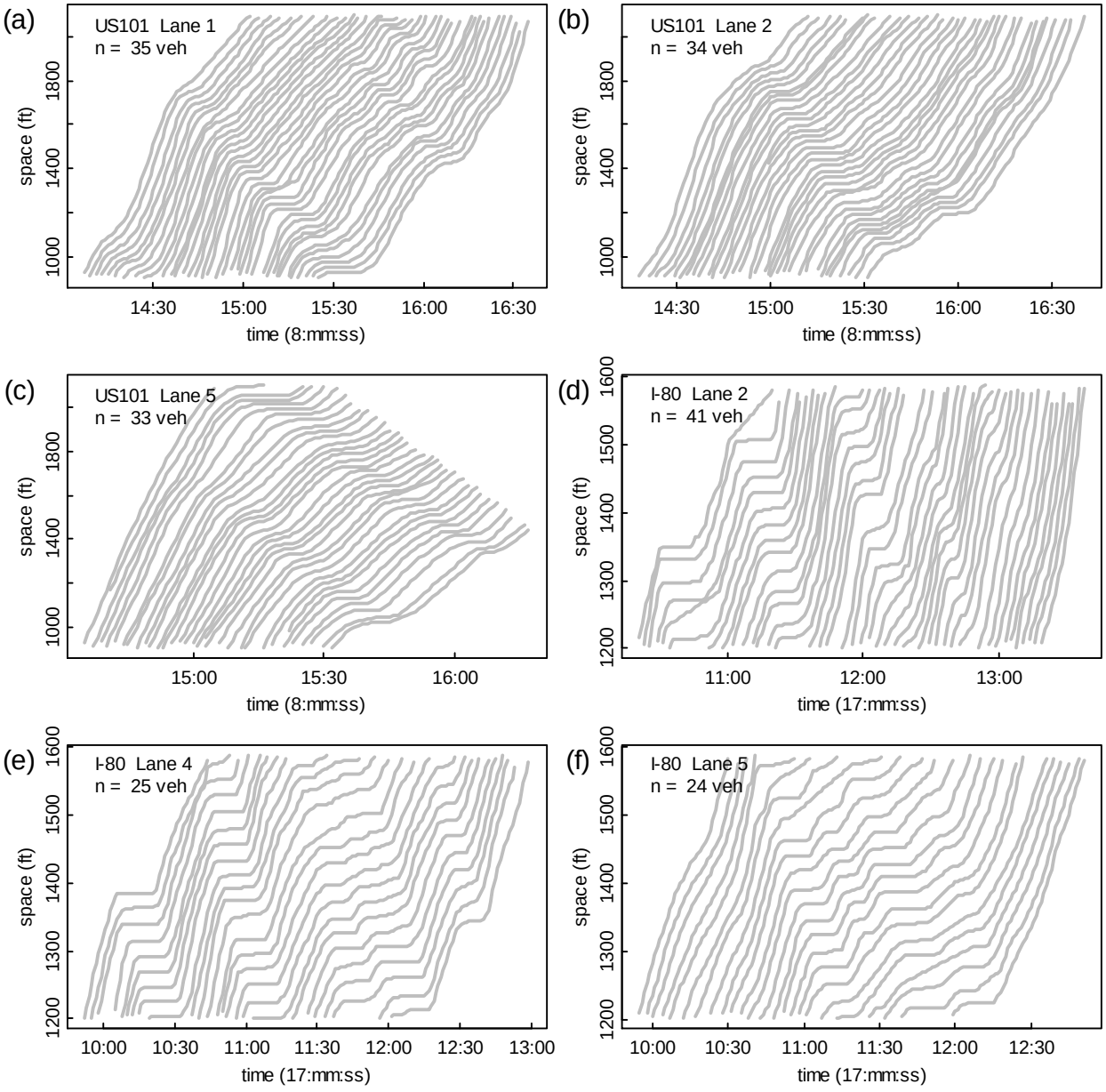

Figure 35 NGSIM vehicle trajectory

The algorithm is then applied to the data to find X_s and X_g. Prior to determining X_s and X_g the trajectories are extracted by their ladders or shockwaves, as shown in Figure 36. The resulting X_s and X_g are illustrated in Figure 37.

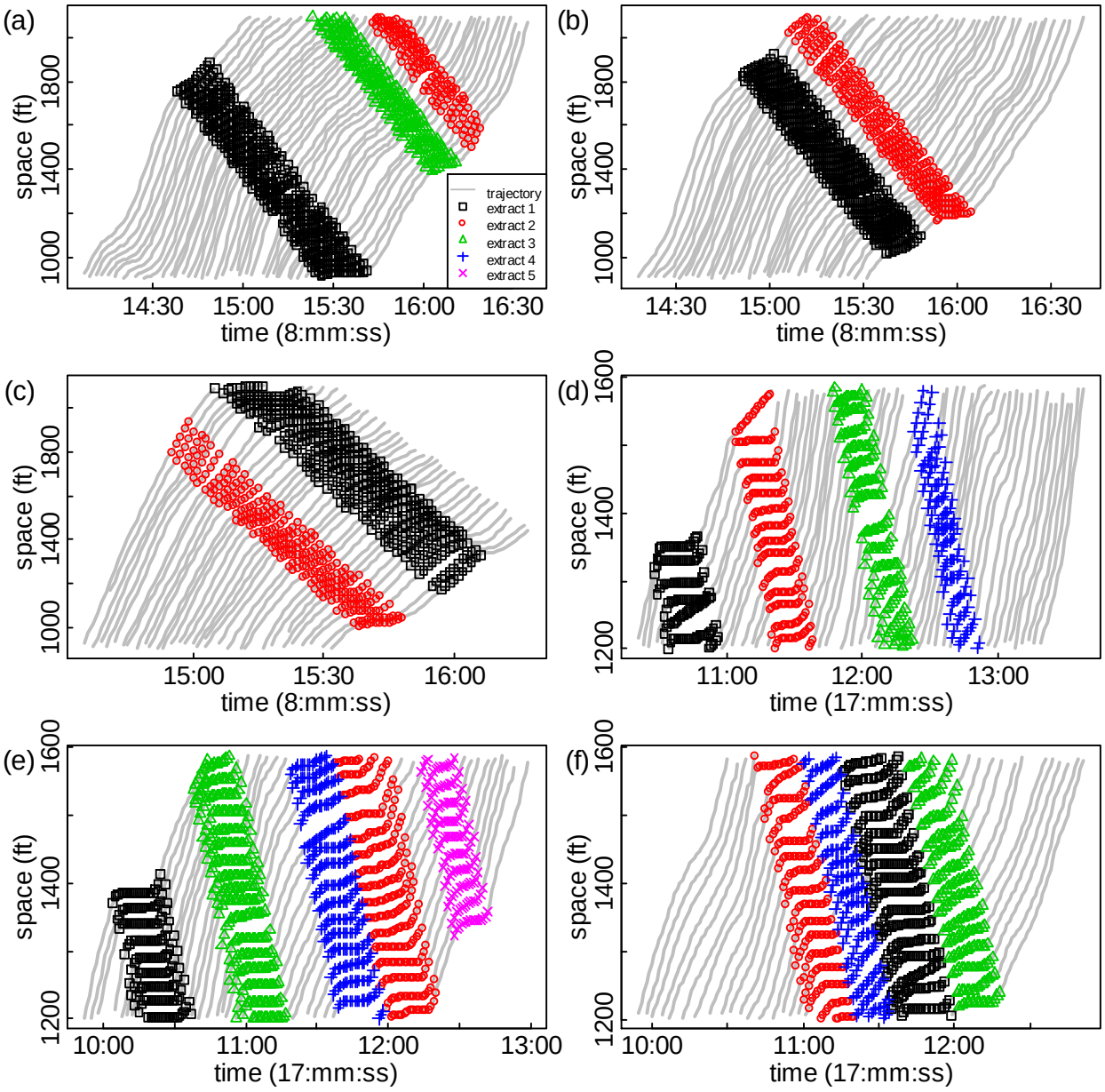

Figure 36 Extracted trajectory by ladder (a-c) US101 lane 1, 2, 5 and (d-f) I-80 lane 2, 4, 5

Figure 36 shows that the trajectories are successfully extracted by their ladders. The extracted points are label accordingly. For US101 lane 2 and 5, the algorithm identified two ladders while lane 1 has three ladders. I-80 lane 2 and 5 have four ladders and lane 4 has five ladders. From

these extracted points, X_s and X_g are then determined, as illustrated in Figure 37. In this figure, the red circles are X_s and green triangles are X_g. It can be observed that even though the points are extracted for each ladder, it doesn't ensure that X_s and X_g exist for that particular trajectory. The ladder is useful identifying the grouping and the possible links of the PV. From these points, ΔX_g are calculated for each η. From here on ΔX_g will be referred to as ΔX.

As stated in the Methodology chapter, after calculating ΔX there exist the overlap and duplicate conditions. These conditions contribute to the correlation of ΔX between η. Figure 38 and Figure 39 illustrate the overlap and duplicate conditions, respectively. In the overlap condition, vehicle sequences are used more than once. To solve this issue, only consecutive vehicle sequences are considered. In duplicate condition, the vehicle sequences are located in multiple ladders. The solution is to simply remove the duplicates.

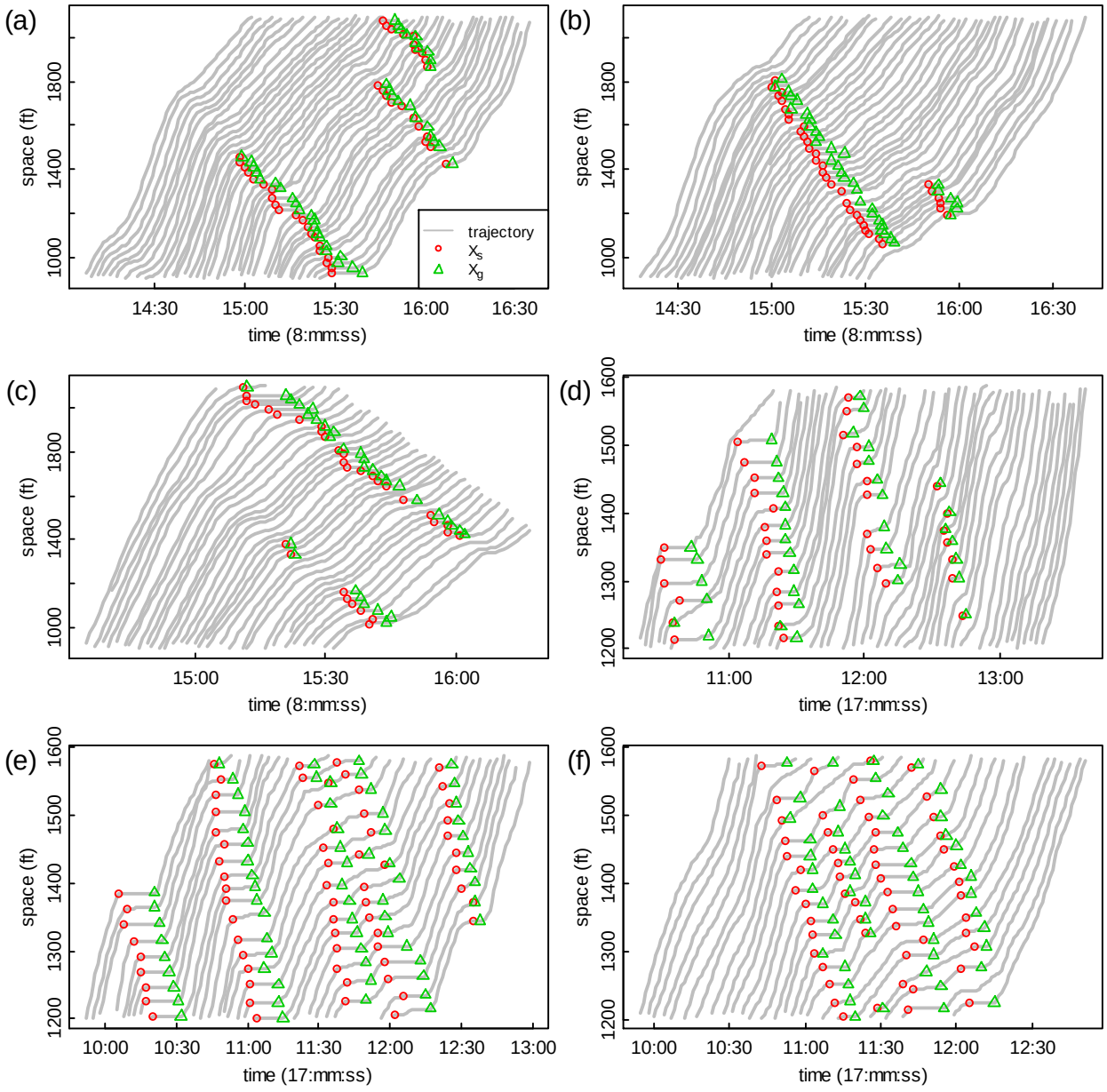

Figure 37 Xs and Xg (a-c) US101 lane 1, 2, 5 and (d-f) I-80 lane 2, 4, 5

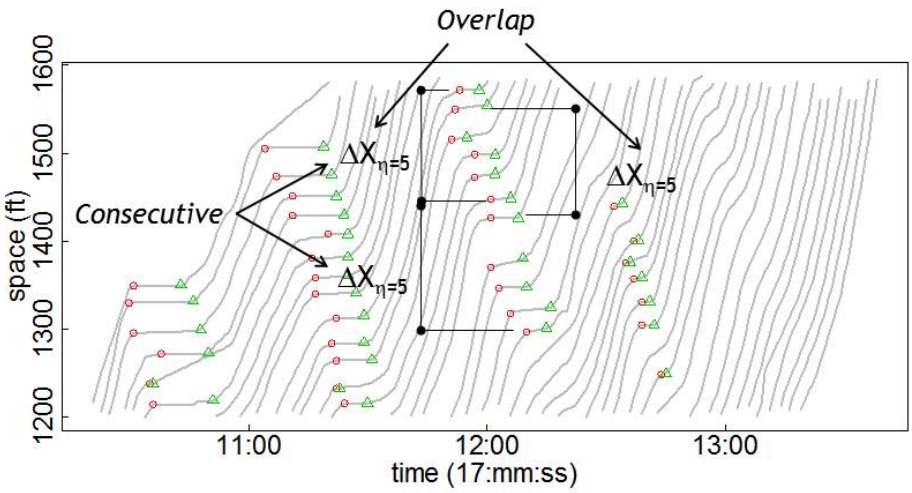

Figure 38 Overlap condition

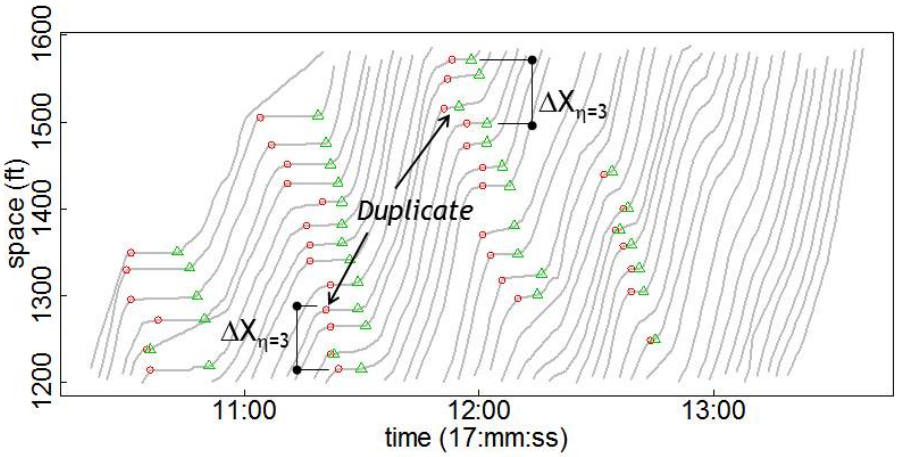

Figure 39 Duplicate condition

Let's assume for a moment that the overlap and duplicate conditions were not removed from the data set. It is expected that both conditions would contribute to the correlation of ΔX for each η. To show this correlation ΔX is plotted for each η, as shown in Figure 40. In this figure, it

can be observed that the correlation of ΔX becomes stronger as η increases. It can be observed that, as η increases the points become less scatter and start to form a 45 degree line. This would then affect the pdf and ε which are used to predict η. It is also noticed that the number of samples decreases as η increases. This can be confirmed by the number of observations for each η. In this figure, the maximum η is set to be at ten.

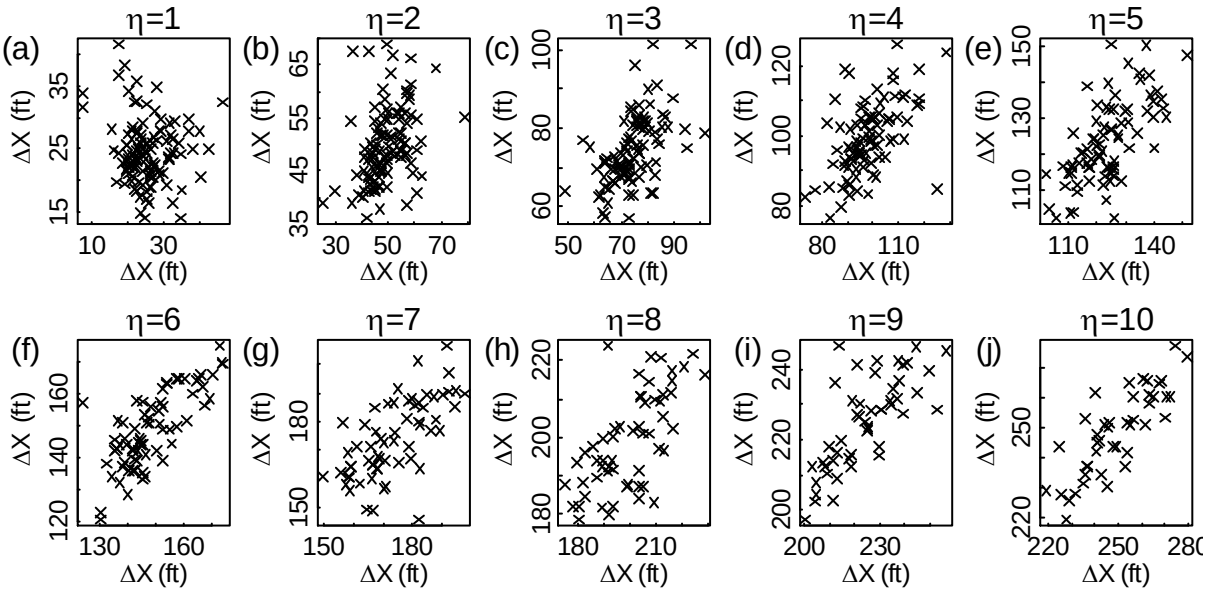

Figure 40 ΔX with overlap and duplicate

After removing the overlaps and duplicates, ΔX is then transformed to g. Figure 41 is a plot of g for each η. It can be observed that g remains independent even as η increases as the points remain scattered. However, by removing the overlap and duplicate, the number of samples is drastically reduced. While removing overlaps and duplicates helps with the characteristic of being independent, it comes at a cost of reduced sample size.

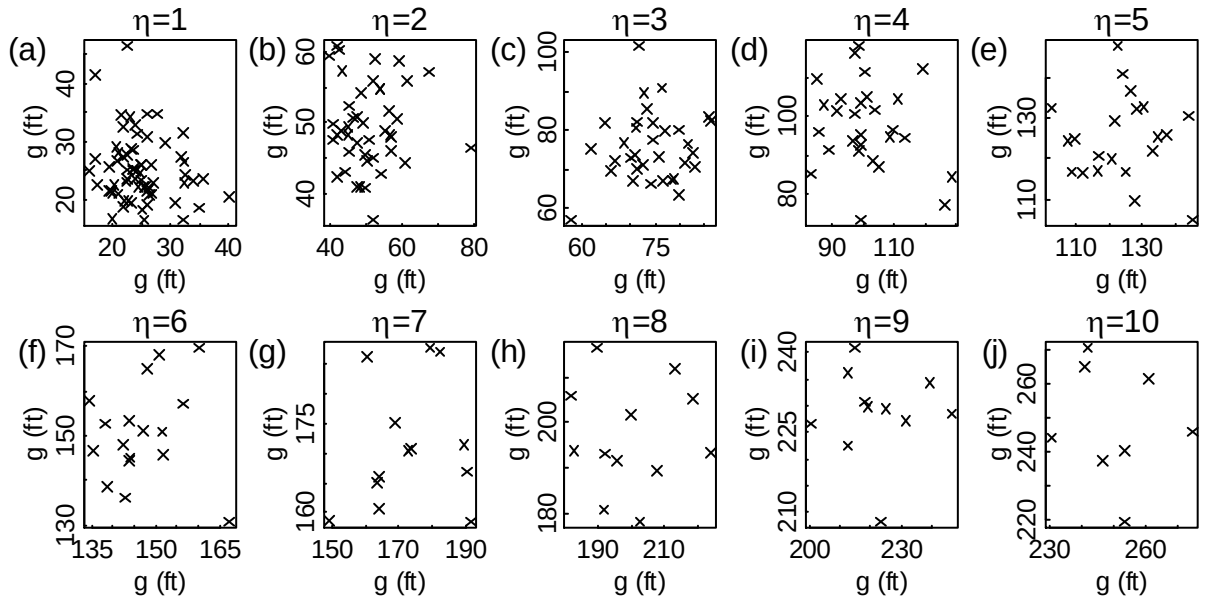

Figure 41 Consecutive and no duplicate of g

Table 4 lists the mean, variance and number of samples of ΔX and g for every η. The numbers in this table are free from overlap and duplicate conditions. As mentioned earlier, as η increases, the number of samples decreases, which makes the training model to be less reliable as η increases. To counter the number of samples issue, a linear regression line is fitted through the data for η from 1 to 15. The results from the regression are listed on the right most column of Table 4.

Figure 42 shows the relationship between the mean and variance of ΔX and g for every η. This figure is essentially a graph form of Table 4. Due to limited number of samples, the plot is only for η from to 1 to 15. The mean for both ΔX and g increases as η increases. It also shows that the mean is linear with respect to η. In general the variance for both ΔX and g also increases as η increases

Table 4 Mean, variance and number of samples of ΔX and g

η	$\mu_{\Delta X}$ (ft)	$\sigma^2_{\Delta X}$ (ft^2)	μ_g (ft)	σ^2_g (ft^2)	Num. Samples	Linear fit σ^2_g (ft^2)
1	25	28	11	25	133	16
2	50	47	21	49	87	41
3	75	71	31	67	67	65
4	99	125	41	125	56	90
5	124	118	51	126	43	115
6	149	96	61	122	34	140
7	172	133	70	145	26	165
8	199	151	81	177	25	190
9	225	127	94	149	22	214
10	249	206	102	229	17	239
11	274	180	114	278	14	264
12	298	177	123	294	15	289
13	323	220	136	306	12	314
14	348	291	144	388	14	339
15	371	193	154	468	7	363
16	403	444	178	686	4	388
17	437	105	200	175	3	413
18	460	149	210	171	3	438
19	481	629	214	410	3	463
20	503	620	222	422	3	487
21	535	493	241	406	3	512
22	558	2178	251	1770	2	537
23	580	2364	260	1907	2	562
24	615	1642	281	1026	2	587
25	638	1345	291	856	2	612
26	659	1356	299	864	2	636
27	682	1081	304	703	2	661
28	701	1320	309	736	2	686

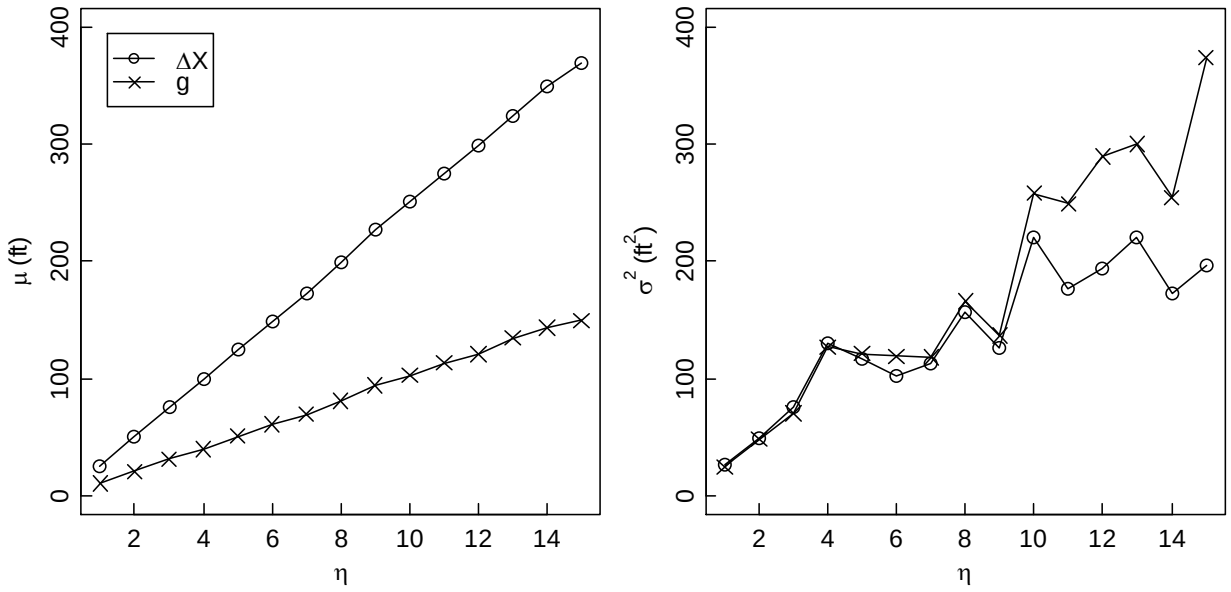

Figure 42 (Left) Mean and (Right) variance of ΔX and g

Figure 43 and Figure 44 are the histogram and empirical pdf of ΔX and g, respectively. They are plotted for each η. When $\eta = 1$ the pdf appears to follow a lognormal distribution. For η greater than 2 the pdf is less obvious. To determine the pdf for each η, the Cullen and Frey (1999) method is employed, as shown in Figure 45. This method relies on the kutosis and skewness of the data. Kutosis is a measure of symmetry and skewness measures the weight of the tail.

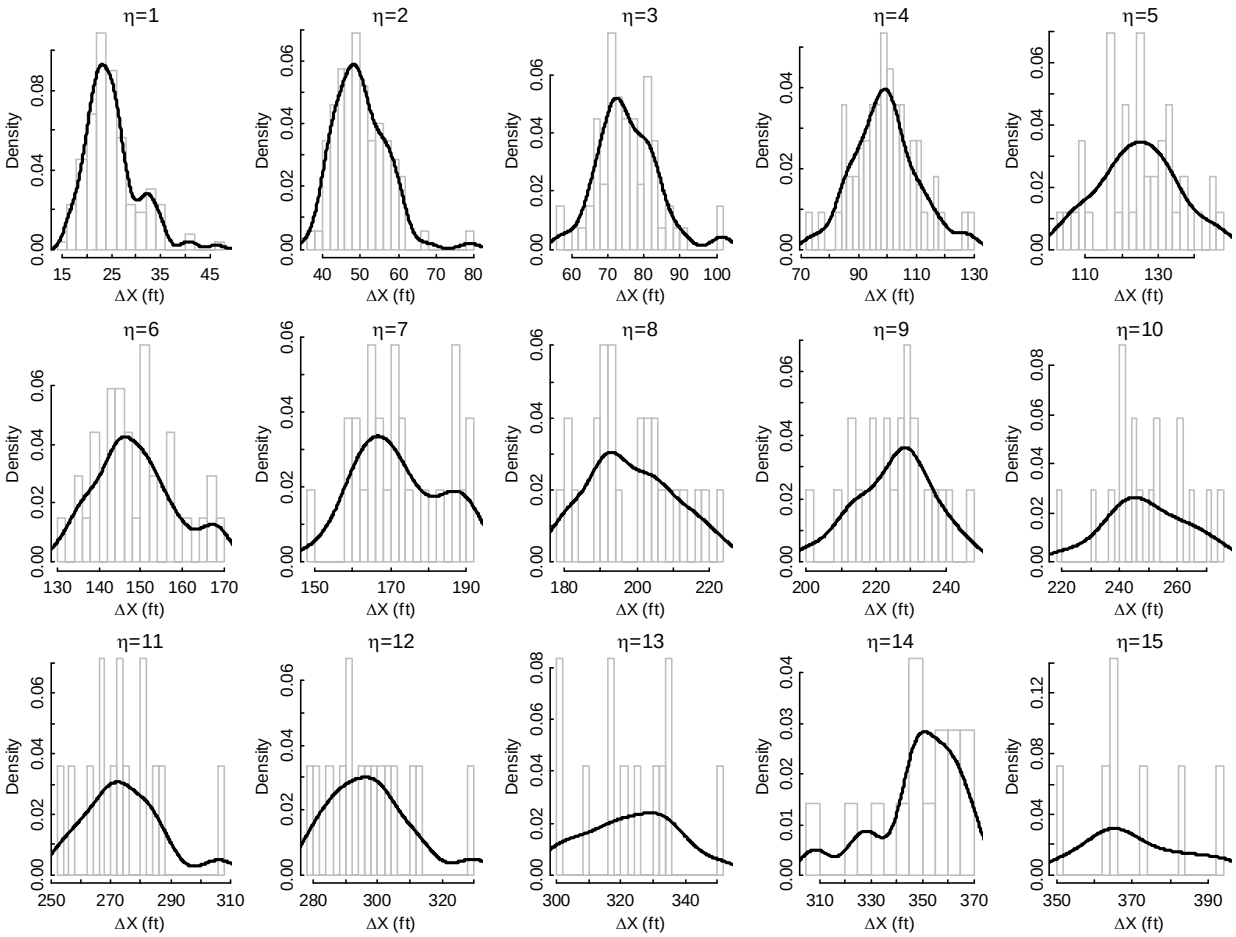

Figure 43 Empirical pdf of ΔX

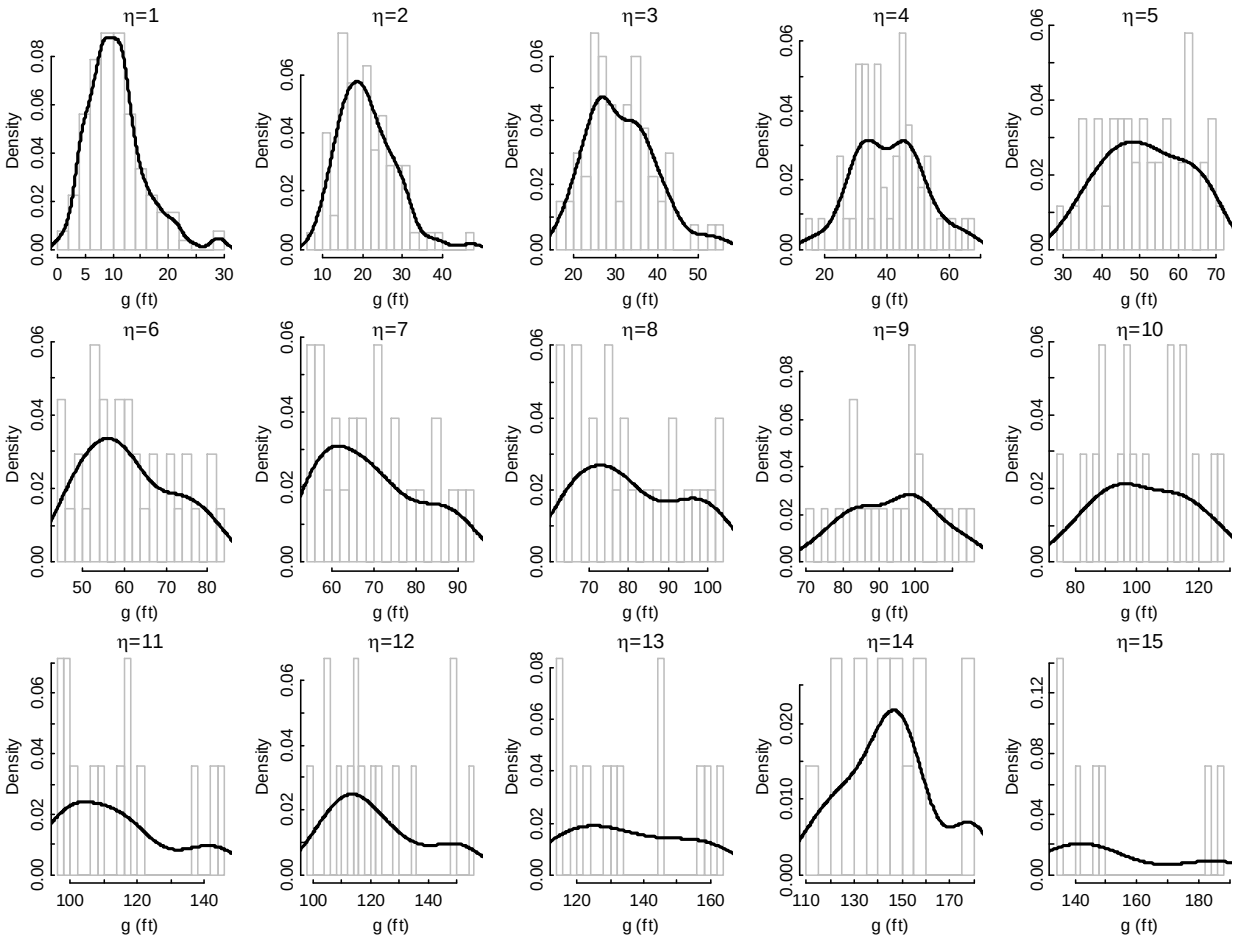

Figure 44 Empirical pdf of g

In Figure 45, which is for $g_{\eta=1}$, the circle is the observed value. The triangle, star and plus are the designations for uniform, normal and logistic distributions, respectively. The small and large dashed lines are the lognormal and gamma distributions, respectively. Beta distribution is the shaded area. This method is an approximation to the type of distribution that g could follow. It does not mean that g will exactly follow a specific distribution. It is a simple test on the possible distributions of the observation.

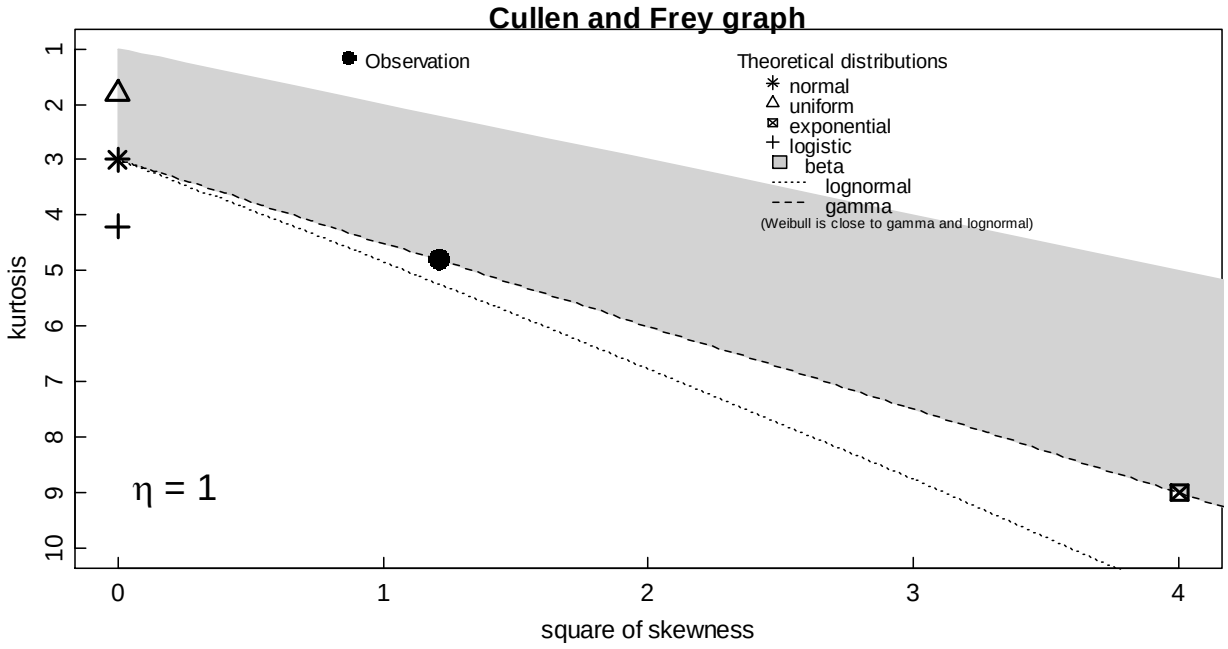

Figure 45 Types of distribution for pdf of g for $\eta = 1$

For $\eta = 1$, it is observed that the pdf of g is located on top of the line signifying a gamma distribution. The pdf of g is also located near the lognormal distribution line. Regardless of the actual distribution, this method shows that the pdf of $g_{\eta=1}$ is heavy on the tail area. With this characteristic, the pdf of g could follow gamma, lognormal or Weibull distribution. Analyses for possible distributions for $\eta = 2$ through 10 is shown in Figure 46.

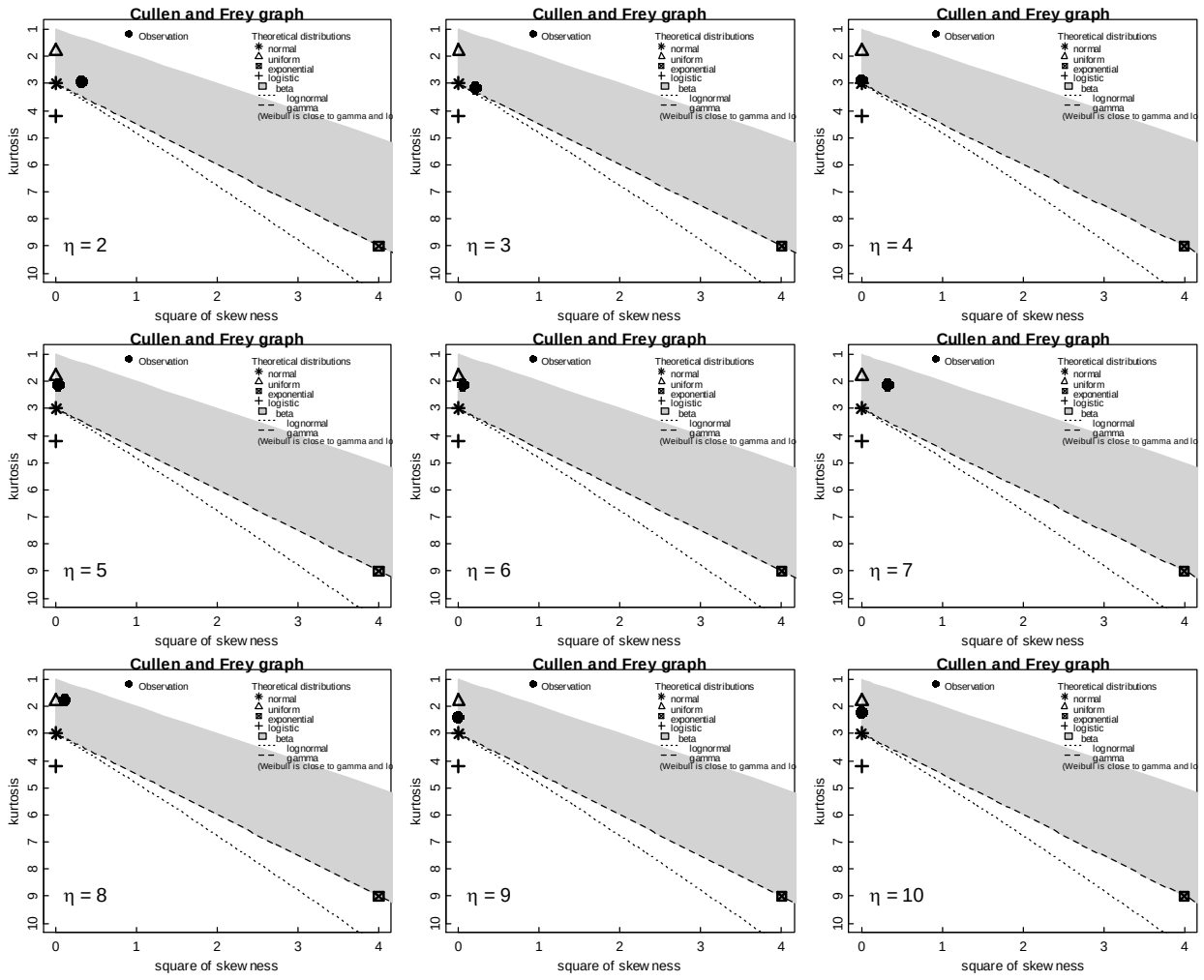

Figure 46 Types of distribution for pdf of g for η = 2, 3, … , 10

In Figure 46, the observed values shifted to the left of the plots. The observed values are located near the points identified as normal and uniform distributions. After analyzing the results for all η, it is decided that pdf of $g_{\eta=1}$ follows lognormal distribution. When η is greater than 1, pdf of $g_{\eta>1}$follows normal distribution.

This concludes the chapter on data. The next chapter is a discussion on the results of the proposed methodologies.

CHAPTER 5

5 RESULTS

This dissertation relies on several different methodologies for estimating traffic flow. The results of each methodology will be discussed separately, beginning with the FD followed by the shockwave then the car-following approach.

5.1 FUNDAMENTAL DIAGRAM APPROACH

The basics of the FD approach is that $\hat{q}$ can be calculated given any u, which is PV speed. In this approach, u is aggregated for every 5 minutes. $\hat{q}$ are then aggregated every 5, 10 and 15 minutes. Figure 47 illustrates $\hat{q}$ for the different FDs at different aggregation intervals. In this figure, the grey, red, black, green and blue lines represent q and $\hat{q}$ from loop detector, Greenshield, Van Aerde, Underwood and Northwestern, respectively.

It is noticed that the deviations of $\hat{q}$ from q are lower when flow is at capacity or close to capacity while it is higher when traffic is mild. Moreover, it is observed that the Greenshields, Underwood and Northwestern tend to underestimate $\hat{q}$ when flow is near capacity. Overall, Van Aerde provides a reasonable $\hat{q}$ during periods of high flow with slight under estimation during medium flow. It is observed that there is less fluctuation as the aggregation increases from 5 to 15 minutes.

Estimation errors for each set of time aggregation and FD are examined using different performance indicators: PE, MAPE and RMSE. Figure 48 shows a summary of the magnitude of the deviations of $\hat{q}$ from q in terms of PE for each FD and aggregation interval. The figure

reveals that the magnitudes of the errors from Greenshields, Underwood and Northwestern are high and contain a significant amount of outliers. Moreover, the errors from Greenshields and Underwood show no significant improvements with the increase in aggregation interval while Northwestern and Van Aerde show reductions in estimation errors with an increase in aggregation intervals. On the other hand, the size of boxes and whiskers of PE for Van Aerde are smaller than the other three FD models which suggest that the variability in the errors of $\hat{q}$ is smaller.

Table 5 lists the MAPE, RMSE, average PE and standard deviation of PE for each FD and aggregation interval. Overall, Van Aerde performs the best in terms of having the smallest magnitude and variation of errors in terms of MAPE, RMSE and average and standard deviation of PE.

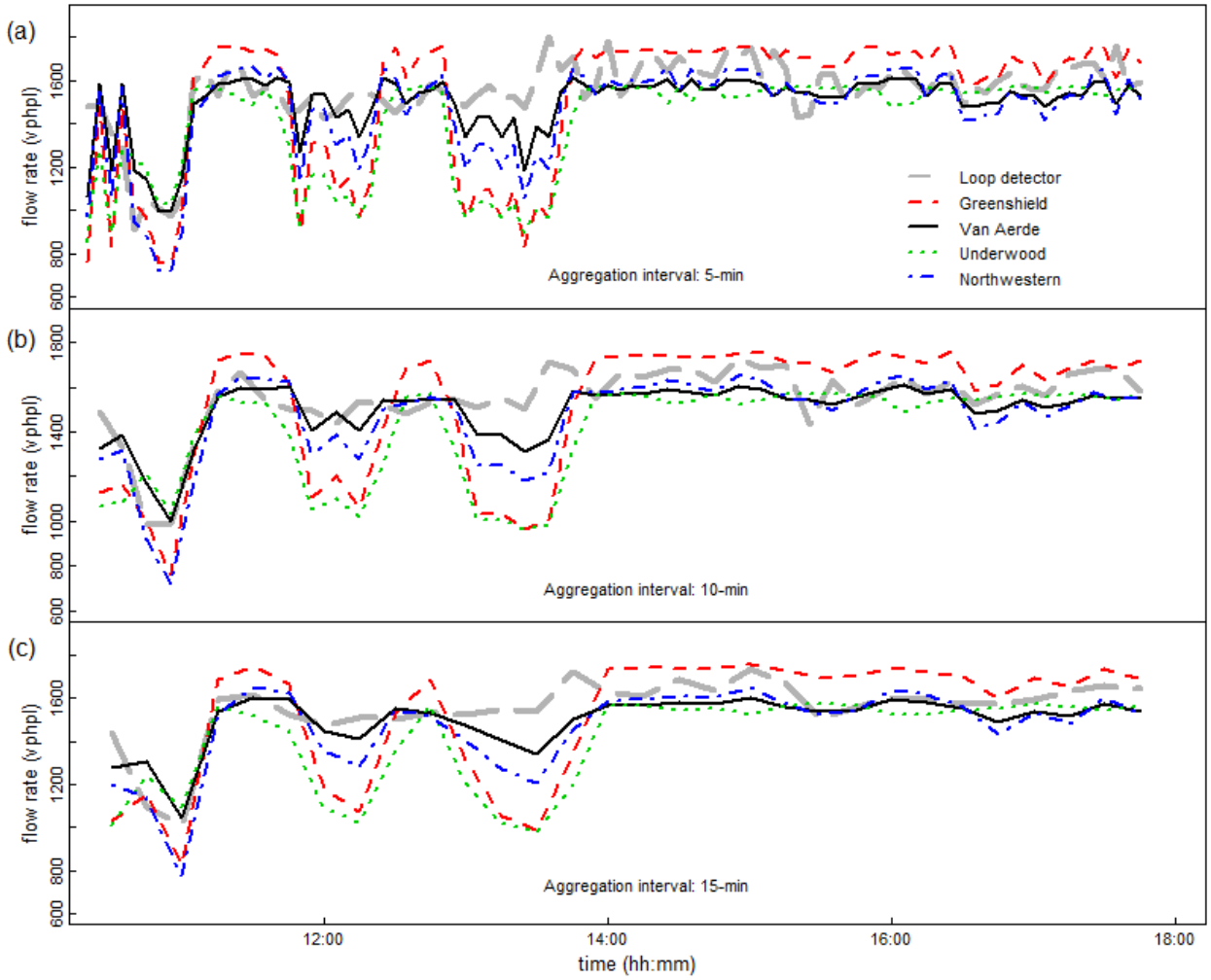

Figure 47 Flow estimation from different fundamental diagrams and aggregation intervals

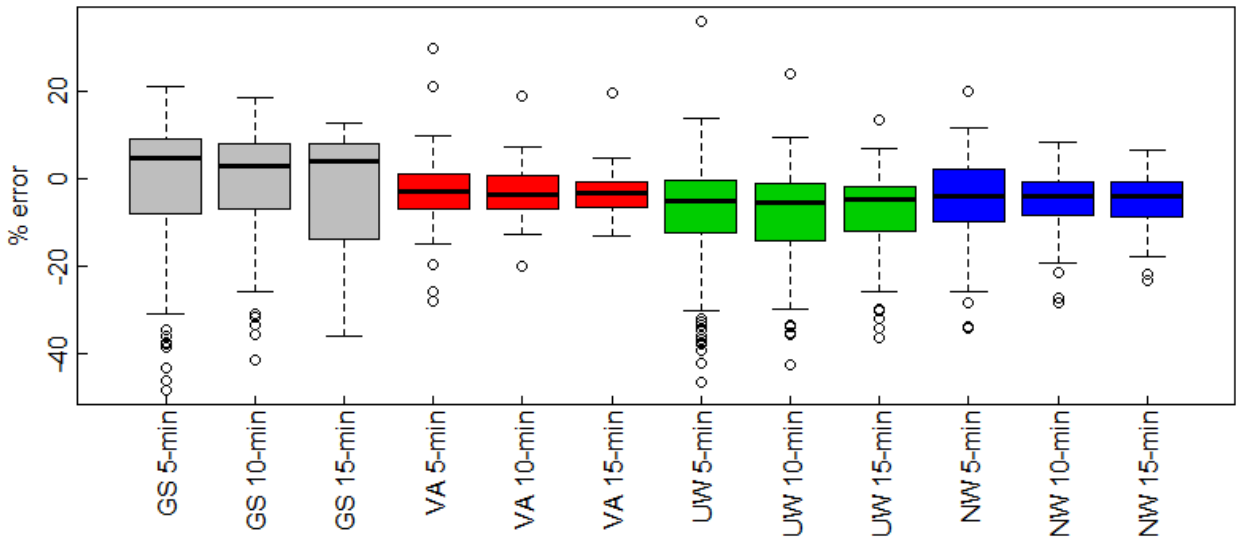

Figure 48 Distribution of percentage error for different fundamental diagrams and aggregation intervals

Table 5 Summary of errors for different fundamental diagrams and aggregation intervals

FD models	Aggregation interval	MAPE (abs %)	RMSE (vphpl)	PE Avg. Error	PE Std. Dev.
Greenshields	5-min	12.5	189	-2.1	17.1
	10-min	11.1	169	-2.2	15.2
	15-min	11.1	168	-2.2	14.7
Underwood	5-min	11.7	178	-8.9	14.6
	10-min	11.3	174	-9.0	13.5
	15-min	10.9	167	-9.0	12.9
Northwestern	5-min	8.7	130	-5.4	10.4
	10-min	7.1	107	-5.5	8.2
	15-min	6.8	103	-5.5	7.7
Van Aerde	5-min	6.4	98	-2.9	8.1
	10-min	5.3	83	-3.0	6.2
	15-min	5.2	79	-3.0	6.2

5.2 SHOCKWAVE APPROACH

To estimate $\hat{q}_f$ using the shockwave approach, the values of w, u_j, u_f and $\hat{q}_j$ are needed. From PV trajectory and *k*-means clustering, values of w, u_j and u_f can be determined. The fourth variable which is $\hat{q}_j$ requires additional computation since it is a function of u_B which is the breakpoint speed.

Prior to estimating $\hat{q}_j$ a proper u_B value must be selected. In this dissertation ω is calculated for every twentieth PV with u_B search values ranges from thirty to fifty mph. u_B is selected when Z is minimum. The results of u_B and Z are shown in Table 6. From this table $u_B = 35$ resulted in the lowest Z at 0.442. Using Northwestern congested regime FD, when $u_B = 35$ mph, $\hat{q}_j = 912$ vphpl. The resulting $\hat{q}_j$ will in turn be used to calculate $\hat{q}_f$.

Figure 49 illustrates the grouping for every 20th PV. In this figure the circles are the transition points (excluding the outliers), dashed lines are trajectories of the first, every 20th and last PV and solid lines are the shockwaves. Note that the number of PV for the last group is less than twenty. The selection of every 20th PV is based upon the vehicles entry time into the network.

Table 6 Results of u_B and Z

u_B	Z
30	0.470
31	0.478
32	0.476
33	0.462
34	0.479
35	0.442
36	0.466
37	0.475
38	0.517
39	0.515
40	0.502
41	0.503
42	0.487
43	0.501
44	0.533
45	0.536
46	0.549
47	0.628
48	0.588
49	0.572
50	0.577

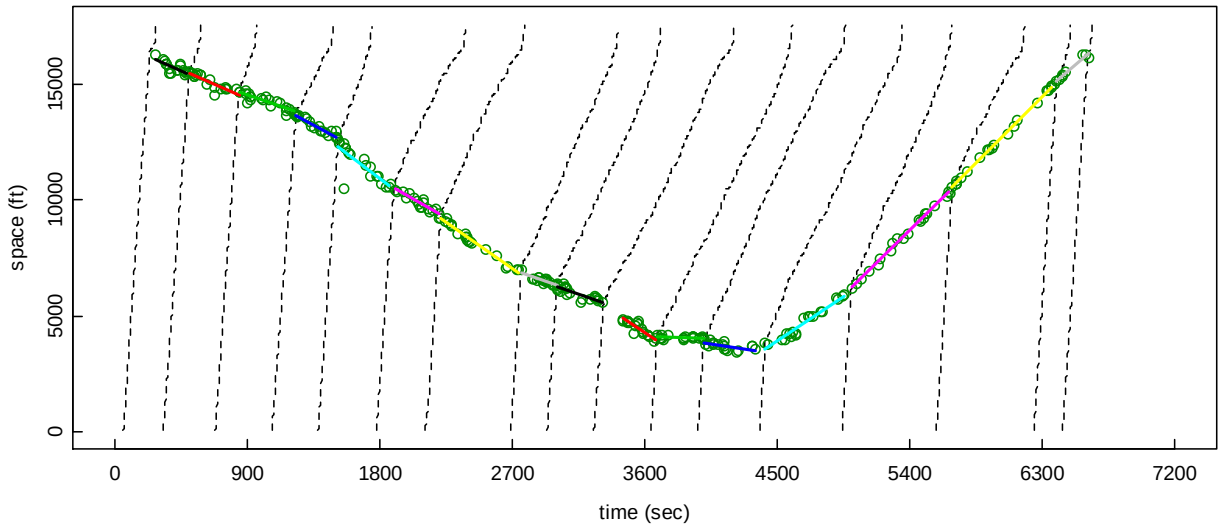

Figure 49 Probe vehicle trajectory for every 20th vehicle, transition points and shockwave

Table 7 summarizes the results of $\hat{q}_f$ and $\hat{q}_j$ when u_B is equal to 35. For comparison purpose, the same table summarizes $\hat{q}_f$ and $\hat{q}_j$ when u_B is equal to 40 and 45. $\hat{q}_f$ and $\hat{q}_j$ are compared to q_f and q_j with the difference calculated in terms of percent error. The formula for percent error is written in the table. Note that u_B=40 is the value proposed by the Northwestern congested regime FD.

Table 7 Summary of results for $\hat{q}_f$ and $\hat{q}_j$

PV Group	Actual flow q_f (vphpl)	$\hat{q}_j$ = 912 (u_B = 35) PE_j = -21%		$\hat{q}_j$ = 1105 (u_B = 40) PE_j = 0%		$\hat{q}_j$ = 1303 (u_B = 45) PE_j = 12%	
		$\hat{q}_f$ (vphpl)	PE_f	$\hat{q}_f$ (vphpl)	PE_f	$\hat{q}_f$ (vphpl)	PE_f
1	1385	1074	-22	1298	-6	1526	10
2	1337	1066	-20	1288	-4	1515	13
3	1337	1030	-23	1246	-7	1466	10
4	1382	1094	-21	1322	-4	1554	12
5	1550	1176	-24	1419	-8	1667	8
6	1451	1106	-24	1335	-8	1569	8
7	1453	1145	-21	1382	-5	1624	12
8	1353	1039	-23	1256	-7	1478	9
9	1328	1023	-23	1236	-7	1455	10
10	1438	1123	-22	1356	-6	1593	11
11	1149	918	-20	1112	-3	1311	14
12	1203	966	-20	1169	-3	1377	14
13	873	652	-25	797	-9	947	8
14	687	546	-21	670	-2	800	16
15	647	530	-18	651	1	779	20
16	702	577	-18	708	1	844	20

Even though the objective function Z indicates that u_B = 35 is the optimum solution, the resulting percent error in flow estimation for both free-flow and congestion is too high. During

free-flow, the smallest percent error PE_f is -18%. In comparison, when u_B=40 and u_B=45, the lowest PE_f is 1% and 8% respectively. The average of the percent error is calculated to be -22% (u_B=35), -5% (u_B=40) and 12% (u_B=45).

In the congested period when u_B=35, $\hat{q}_j$ is calculated to be 912 vphpl. Comparing that to q_j which is 1159 vphpl, the PE_j is calculated to be -21%. When u_B=40 and u_B=45, the PE_j is 0% and 12%, respectively. The results indicate that the $u_B$value of 40 which was proposed by the Northwestern congested regime FD gave the best flow estimation.

5.3 CAR-FOLLOWING APPROACH

In this approach, an algorithm is developed to determine ΔX for each leader and follower pairs, which is then transformed into g. To predict η is to take the shortest path between the lead and last vehicles with ε being the cost function.

To begin the prediction of η a set of PV data are process through the algorithm. The results are the X_s and $X_g$of the PV. An example is shown in Figure 50. In this figure, there are six PV, labeled 1 through 6. The algorithm identified X_s and X_g for PV1 through PV5. The sixth PV did not have any X_s and X_g points, therefore a prediction cannot be made to this PV. The prediction of η will be PV5 with respect to PV1. The leader and last follower are PV1 and PV5, respectively. PV2, PV3 and PV4 are intermediary PV. In reality, there are other trajectories from non-PV located in between the observed trajectories.

There are three different ladders for this group of PV. The first ladder is a one step ladder located to the left of the plot. The step is created from PV1 only. The second ladder is a four

steps ladder located in the middle with the steps developed from PV1, PV2, PV3 and PV4. The last ladder is to the right with the steps developed from PV3, PV4 and PV5.

The link connection is complete for this group. PV1 provides the link between the left and middle ladder. PV3 and PV4 provide the links between the middle and right ladder. This can be observed by the presence of X_s and X_g by each vehicle in the different ladders. PV2 is not connected to the right ladder because its X_s and X_g are located only in the middle ladder.

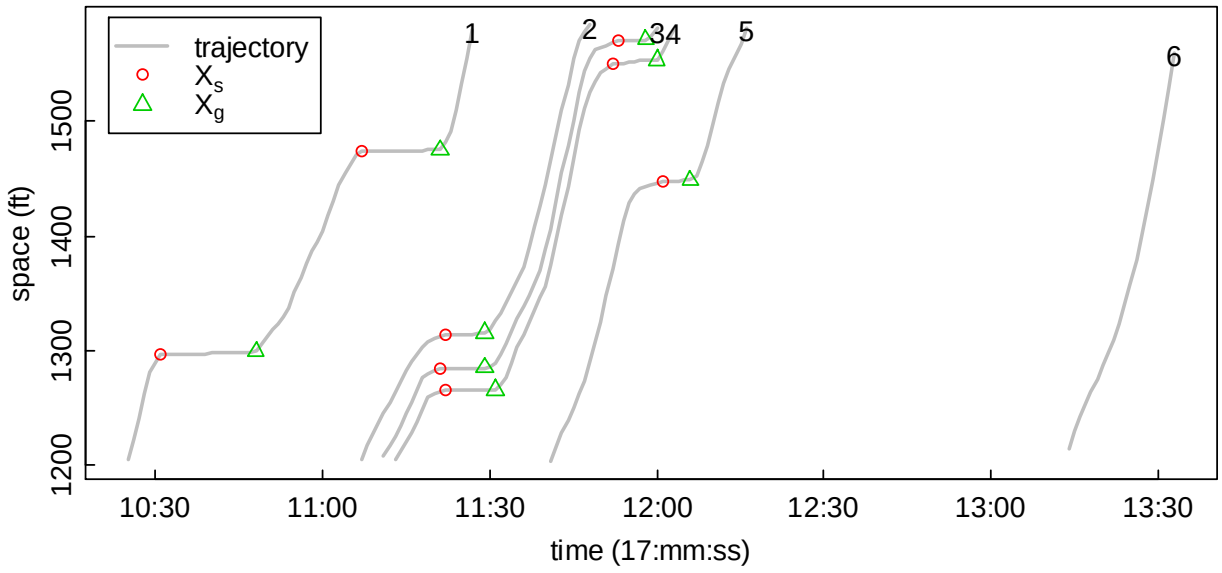

Figure 50 Probe vehicle trajectory, Xs and Xg

For each ladder ΔX_g is calculated for every vehicle pair. For the middle ladder, the vehicle pairs are 1-2, 1-3, 1-4, 2-3, 2-4 and 3-4. For the right ladder, the vehicle pairs are 3-4, 3-5 and 4-5. There is no vehicle pair for the left ladder. ΔX_g for each ladder and for every vehicle pair is

then transformed into g. Based on the pdf parameters developed from the training data and by taking the shortest path, η is predicted from PV5 with respect to PV1 based on g value.

An example of an output from the algorithm is listed in Table 8. This output is not related to the PV in Figure 50. This table lists the pairings of lead and follower PV, actual and predicted η and the ε. In the first line, the pairs are PV1 and PV2. The actual and predicted η are 1 and 2, respectively. The last column on this line is the associated ε with making the prediction. The remainder of the lines is the results for different PV pair. In this example, the prediction of η is from PV1 to PV10. One possible prediction is the path of PV 1-6-10 which resulted in predicted η of 20 (9 + 11). This path is shaded. Another prediction is the path of PV 1-5-6-10 with predicted η of 21. This path has the asterisk next to the PV lead. The actual η fpr PV 1-10 is 19. To determine the optimal path, the shortest path problem is utilized. Utilizing Dijkstra's algorithm, the path with minimum ε will be selected. In the nearest node approach, the links to the nearest PV will be selected, regardless of ε. Resulting in a path of PV 1-2-3-4-5-6-10 with predicted η of 22.

Table 8 Sample of PV pair ε, actual and predicted η

PV lead	PV follow	Actual η	Predicted η	ε
1*	2	1	2	40
1	3	2	3	67
1	4	4	4	119
1	5	5	5	125
1	6	10	9	150
2	3	1	1	21
2	4	3	2	40
2	5	4	3	67
2	6	9	9	150
3	4	2	2	40
3	5	3	3	67
3	6	8	9	150
4	5	1	1	21
4	6	6	6	120
5*	6	5	5	125
6	7	3	3	67
6	8	4	5	125
6	9	8	10	230
6*	10	9	11	279

In the prior examples, the link connection is complete between the source and target PV. There are two other conditions named disjoint and intersect connections. Figure 51 illustrates a disjoint condition. The goal is to predict η for PV 1 and 13. Note that PV14 does not have any X_s and X_g therefore it is excluded from analyses. In this example, the disjoint occurs between PV 9 and 10. This is happening due to X_s and X_g for PV 9 and 10 being located in different ladders. Thus dividing the PV into two groups of PV 1 through 9 and PV 10 through 13. Due to the missing link, no prediction can be made between these two groups. The groups can be connected if X_s and X_g for PV 7, 8 or 9 are located in the same ladder as PV 10, 11, 12 or 13. Another

possibility is that X_s and X_g for PV 10 is located in the same ladder as PV 5, 6, 7, 8 or 9. Due to the missing link, η will be predicted from T for PV 9 and 10.

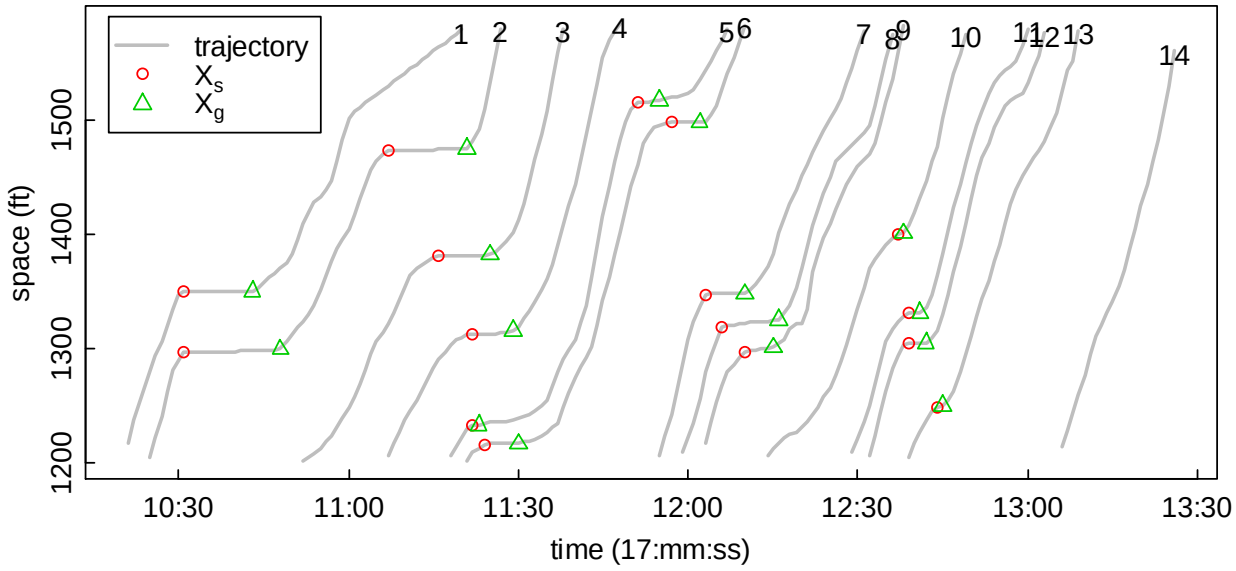

Figure 51 Probe vehicle trajectory, Xs and Xg for disjoint condition

The intersect condition is shown in Figure 52. The goal is to predict η for PV 1 and 13. The intersection occurs at PV 10 where its X_s and X_g are not located in both ladders. Because of this, the PV are divided into groups of PV 1 through 12 and PV 10 through 13. Here the intersect is between PV 10 an 12. Similar to the disjoint condition, η will be predicted from T for PV 10 and 11.

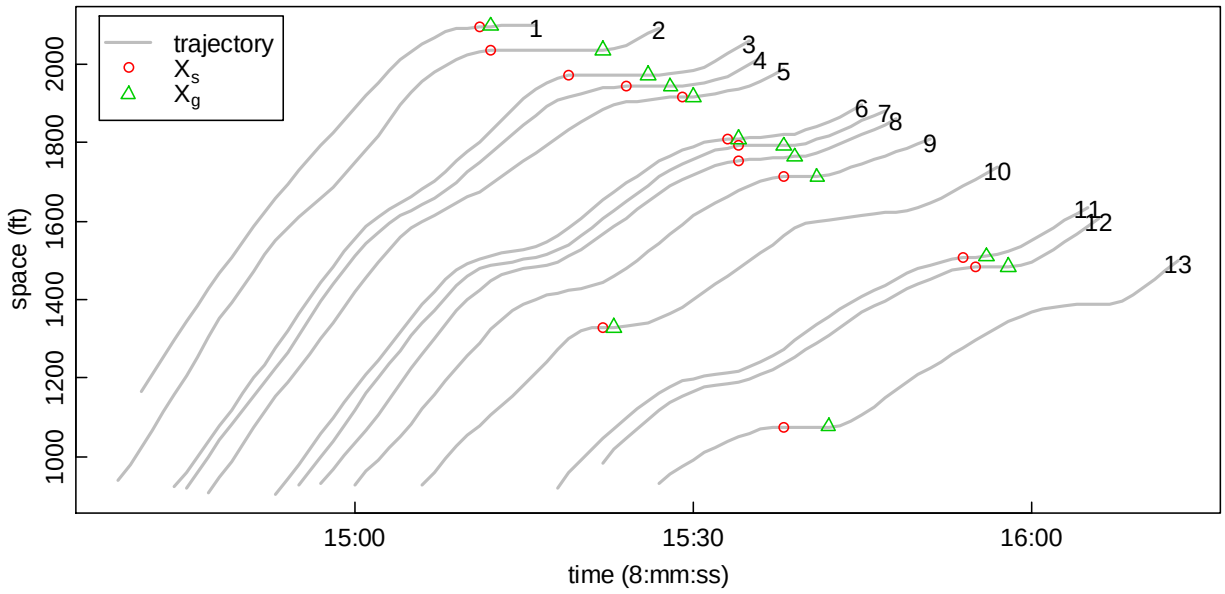

Figure 52 Probe vehicle trajectory, Xs and Xg for intersect condition

If the connection is complete, η is predicted based solely on the variable g. However when there is disjoint or intersect, η will be predicted from a combination of variables g and T. To understand the final prediction of η it is important to distinguish the prediction by the variables g and T. Table 9 is an example results for complete, disjoint and intersect conditions.

In Table 9, when it's a complete condition the prediction is based on variable g only. When its disjoint, the prediction is the sum of all η predicted by variables g and T. In an intersect condition, the prediction from T is deducted from the sum predicted by g. To further understand the prediction made from g and T, Figure 53 compares the difference between the actual and predicted η for the different variables g and T.

Table 9 Example results for complete, disjoint and intersect conditions

Condition	From PV	To PV	Predicted η	Actual η	Variable
Complete	1	5	13	13	g
Disjoint	1	7	17	16	g
	8	9	1	1	g
	7	8	7	7	T
sum			25	24	
Intersect	1	6	12	12	g
	4	9	9	9	g
	4	6	-5	-6	T
sum			16	15	

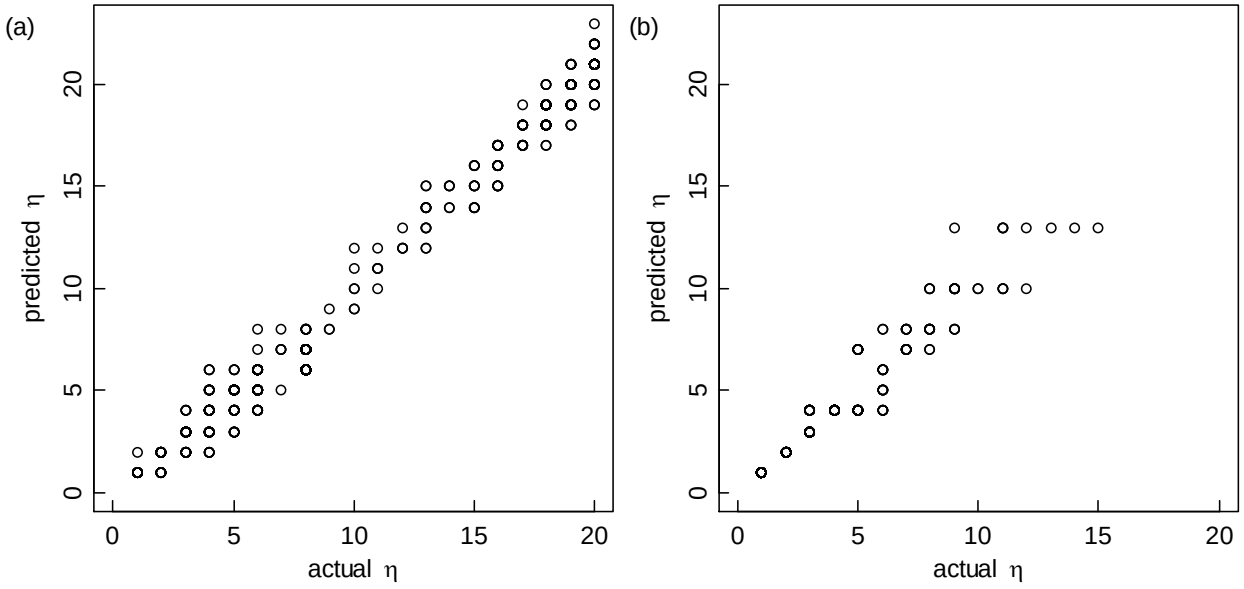

Figure 53 Actual and predicted η using variables (a) g and (b) T

In Figure 53 when the prediction is based on g, the difference between the predicted and actual η ranges from -2 to 2 vehicles. The range is slightly higher which is in between -2 to 3

vehicles when the prediction is based on T. The data for this figure were generated from a specific location. The results would vary as the location changes.

To evaluate the results for the different locations, *k*-fold cross validation is implemented. The configuration of the *k*-fold cross validation is listed in Table 10. There is a total of six different datasets. For each fold, five datasets are selected to be as training data and the remaining dataset will be used as testing. Repeat the process by alternating the test data until all data have been tested.

Table 10 *k*-fold cross validation

Fold	US101 Lane 1	US101 Lane 2	US101 Lane 5	I-80 Lane 2	I-80 Lane 4	I-80 Lane 5
1	Train	Train	Train	Train	Train	Test
2	Train	Train	Train	Train	Test	Train
3	Train	Train	Train	Test	Train	Train
4	Train	Train	Test	Train	Train	Train
5	Train	Test	Train	Train	Train	Train
6	Test	Train	Train	Train	Train	Train

The variance of g for each fold is listed on Table 11. In this table, the results are only for η from 1 through 10. In general, for each fold, the variance increases as η increases. The variances looks similar with the exception of fold 5. For this fold, when η is greater than 5, the rate of increase is smaller when compared to the other folds.

Table 11 Variance of g for each fold

η	1	2	3	4	5	6	7	8	9	10
Fold 1	27	52	62	111	120	132	161	197	164	252
Fold 2	26	52	66	144	133	147	171	191	159	220
Fold 3	21	40	67	119	125	120	148	180	150	230
Fold 4	24	48	70	127	121	119	118	166	137	258
Fold 5	26	49	68	120	128	107	116	137	129	168
Fold 6	27	51	68	130	130	106	158	189	156	245

During testing, PV penetration rate p is varied from 0.05 (5%) to 1 (100%). p is the percentage of PV with respect to the overall vehicle. The PV are randomly selected from a pool of vehicles. When p =1, it means that all vehicles are PV.

It was mentioned earlier that $g_{\eta=1} \sim LN(\mu, \sigma^2)$ and $g_{\eta>1} \sim N(\mu, \sigma^2)$. As comparison, testing will be performed assuming that distribution of ΔX follows normal distribution for all η, $\Delta X \sim N(\mu, \sigma^2)$.

In essence, the testing will be performed for multiple scenarios which are:

- k-fold cross validation (see Table 10)
- Shortest path vs Nearest Node
- Varying p
- Lognormal distribution of g when η = 1 and normal distribution of ΔX for all η

Each test is then repeated 30 times. When the p is low, it is expected that results would vary due to differences in randomly selecting PV. As the p increases, the results will become similar as the PV are the same for each test.

Figure 54 and Figure 55 is the PE of the predicted against actual η for varying p for $g_{\eta=1} \sim LN(\mu, \sigma^2)$ utilizing shortest path and nearest node methods, respectively. The results in Figure 54 is the PE when solved with shortest path under the assumption of $g_{\eta=1} \sim LN(\mu, \sigma^2)$. In this figure, PE decreases as p increases. This can be observed by the reduction of the boxplot size. At p = 1, all vehicles are considered PV. Since the positions of all vehicles are known, it is expected that the PE to be zero or near zero when p = 1. This is true for Figure 54b, d, e and f. In these plots, the PE converges to zero as p increases. Such is not the case for Figure 54a and c where the predicted η is higher than the actual value. This can be explained by the predicted η being higher than actual η. For example, actual η = 1 is predicted as η = 2. Since it is predicted as η = 2 than the ε would be lower compared to η = 1. As a result the shortest path will choose the link with lowest ε resulting in the over estimation of η. The reason that actual η = 1 is predicted as η = 2 is due to vehicle spacing being greater than the historical value. As an example, for η = 1 and η = 2 the mean vehicle spacing is 11 and 21 feet, respectively. However, due to error in data or misidentification of ΔX, it is possible that the vehicle spacing for η = 1 is recognized as η = 2. This error could happen to any η, not just for η = 1 and η = 2.

The nearest node is then applied to predicting η with the results shown in Figure 55. Similar trend can be observed where the PE decreases as p increases. Overall it can be observed that the nearest node is over-estimating η. When the nearest vehicle (or node) is directly behind the lead, the predicted η can be 1 or larger, it cannot be smaller than 1. In this instance when the predicted η is greater than 1, then this value will continue to be used until η is predicted for the last follower. Resulting in over-estimation of η.

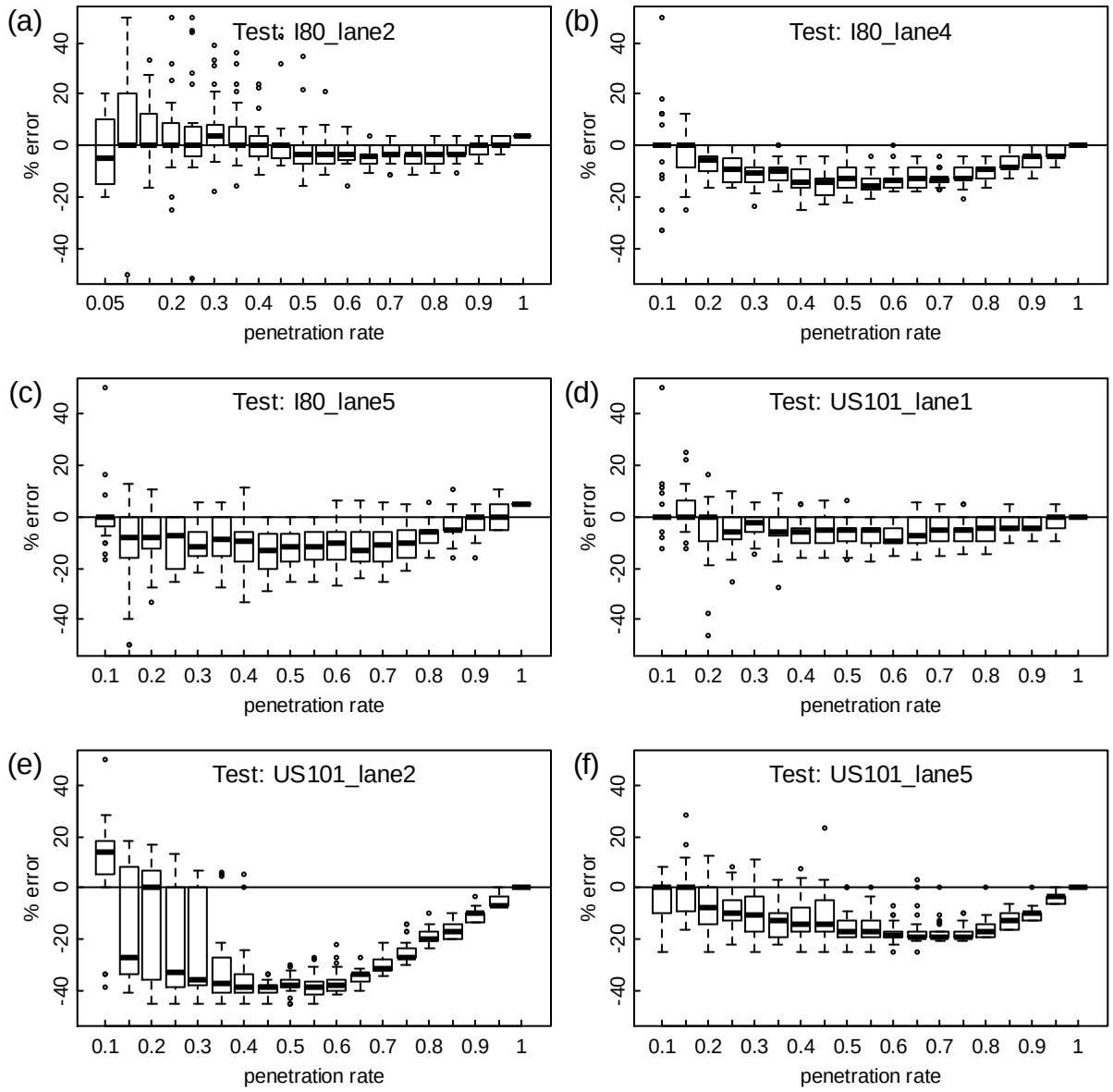

Figure 54 Percent error of η for $g_{\eta=1} \sim LN(\mu, \sigma^2)$ solve with shortest path

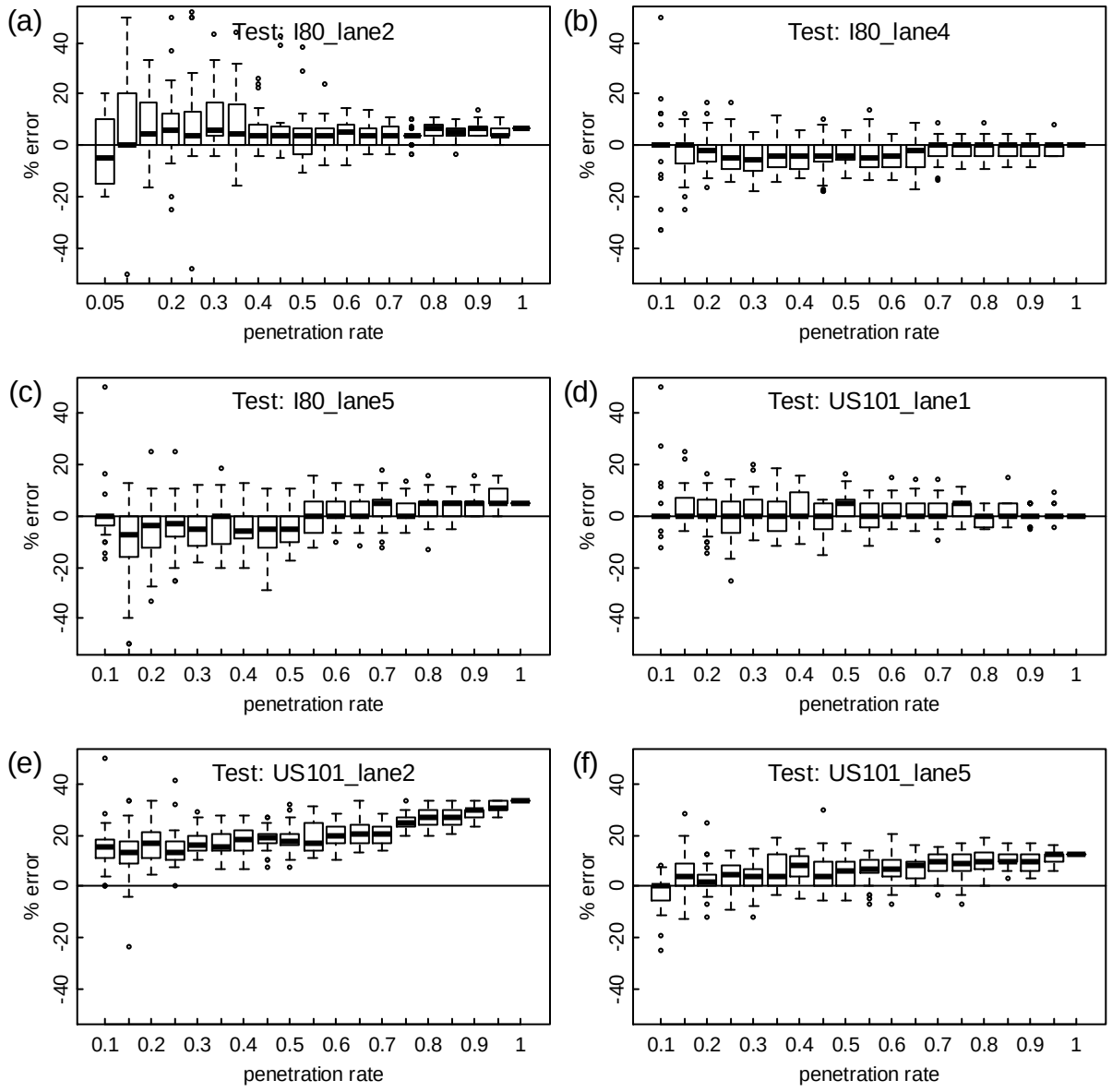

Figure 55 Percent error of η for $g_{\eta=1} \sim LN(\mu, \sigma^2)$ solve with nearest node

Figure 56 and Figure 57 is the PE of the predicted against actual η for varying p for $\Delta X \sim N(\mu, \sigma^2)$ utilizing shortest path and nearest node methods, respectively. Similar to previous figures, PE decreases as p increases. However, the boxplots are smaller when compared

to Figure 54 and Figure 55. This indicates that there is less error at low p. The reasons for over-estimation of η for shortest path and nearest node are the same reasons as described earlier.

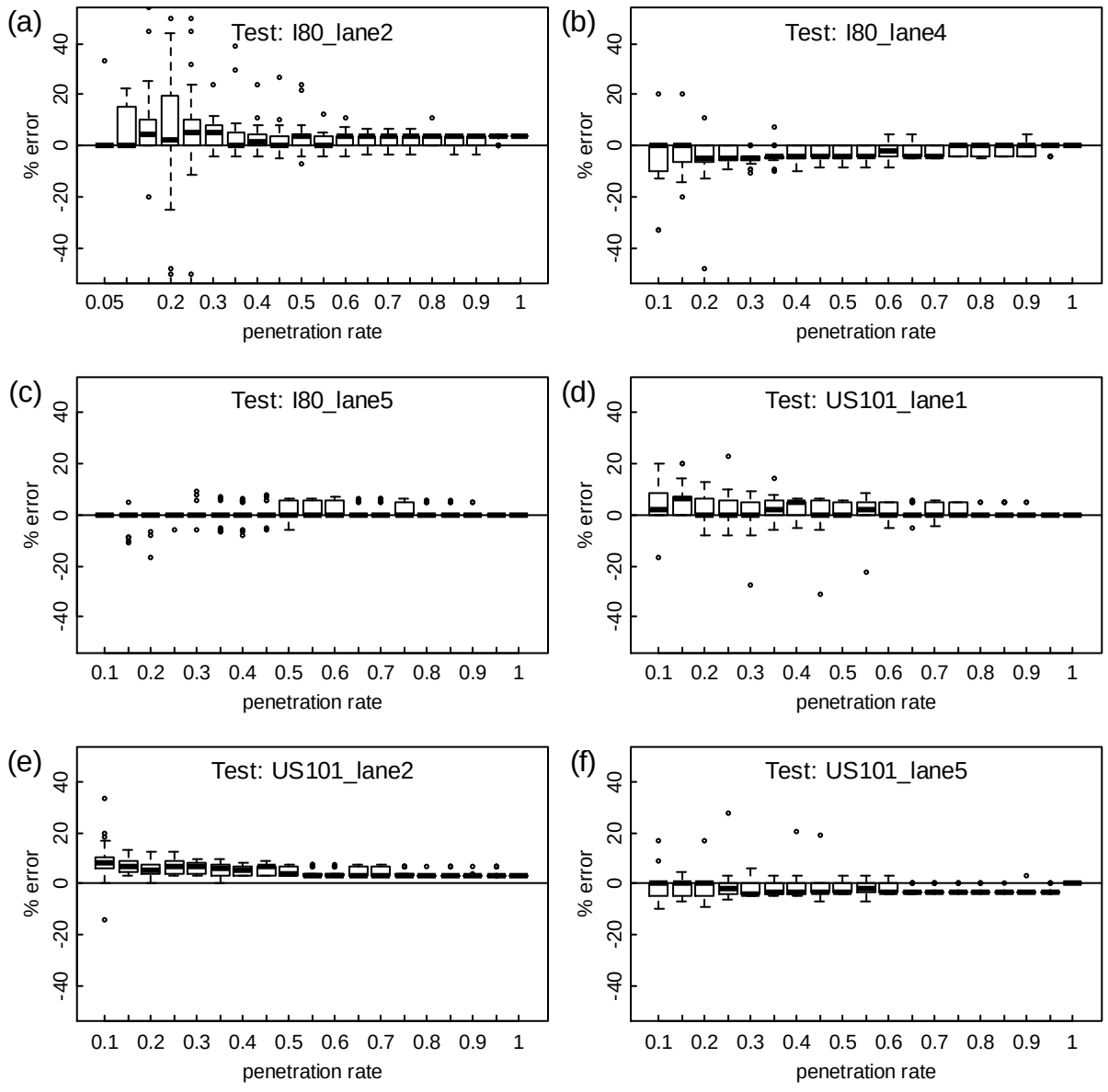

Figure 56 Percent error of η for $\Delta X \sim N(\mu, \sigma^2)$ solve with shortest path

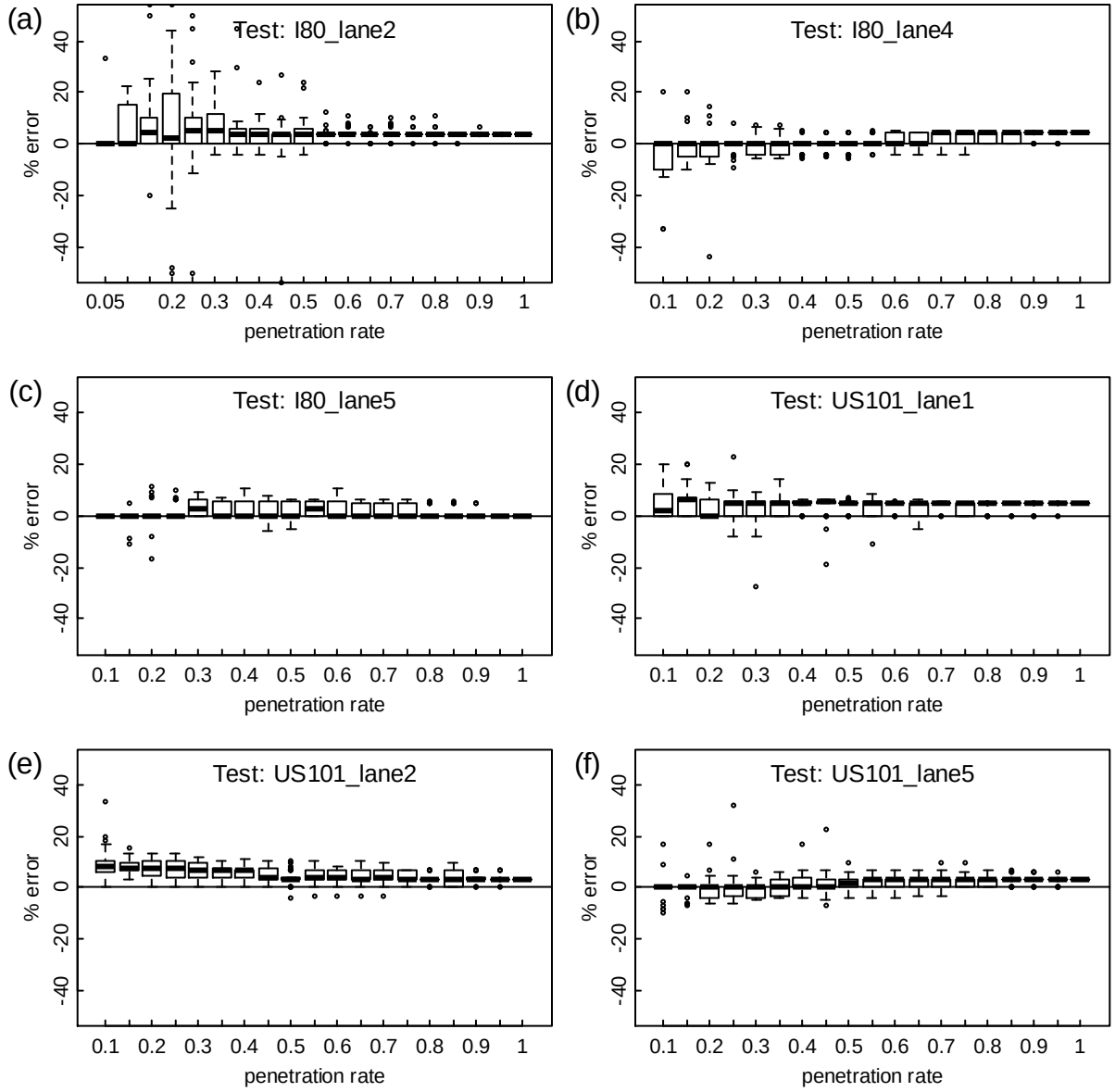

Figure 57 Percent error of η for $\Delta X \sim N(\mu, \sigma^2)$ solve with nearest node

To analyze the results between the different types of variables (g vs ΔX), the MAPE measure will be used. Figure 58 and Figure 59 are the results for the different test data for $g_{\eta=1} \sim LN(\mu, \sigma^2)$ and $\Delta X \sim N(\mu, \sigma^2)$, respectively. In Figure 58a which is solved for $g_{\eta=1} \sim LN(\mu, \sigma^2)$ using shortest path, in general the MAPE decreases as p increases with the

exception of US101 lane 2. Using the nearest node method, the MAPE tends to remain the same with exception of US101 lane 2. The average of all six datasets is labeled in yellow color. On average the MAPE for shortest path decreases as p increases while the MAPE for nearest node remains the same as p increases.

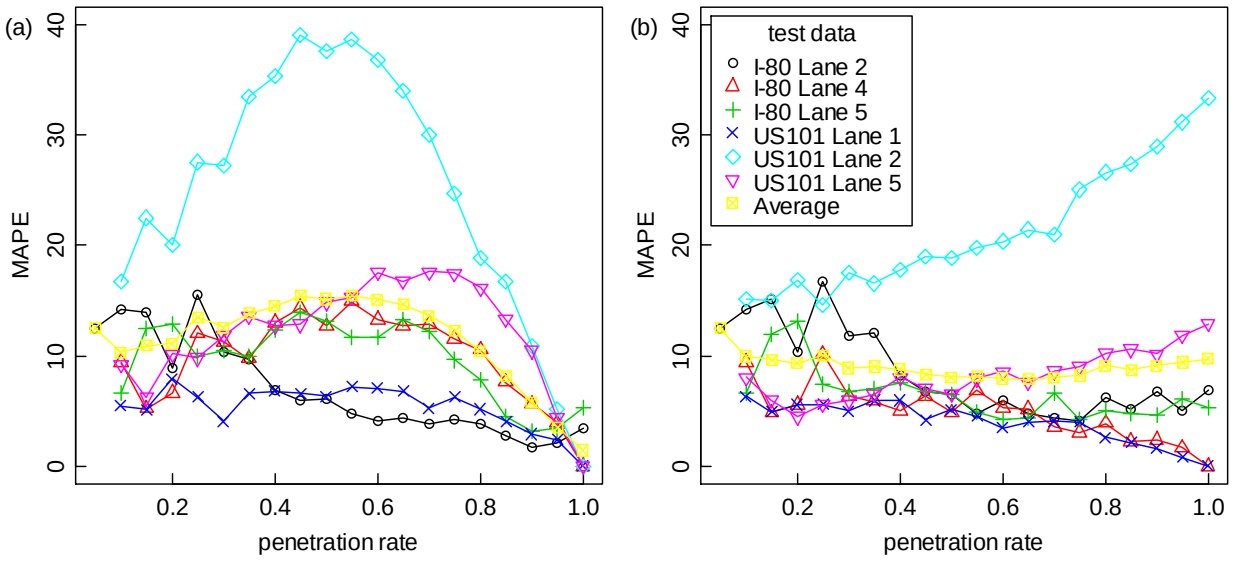

Figure 58 MAPE of η for $g_{\eta=1} \sim LN(\mu, \sigma^2)$ (a) shortest path and (b) nearest node

Figure 59 shows an improved MAPE compared to Figure 58. For comparison the scale of y-axes remains the same as Figure 58. In Figure 59 the MAPE decreases as p increases for both shortest path and nearest node.

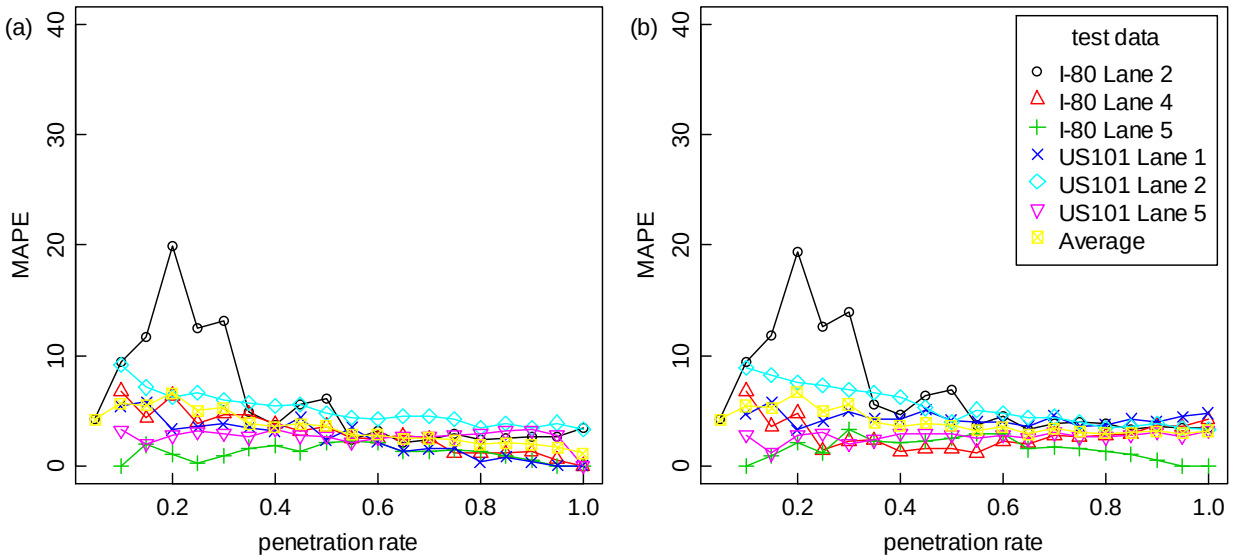

Figure 59 MAPE of η for $\Delta X \sim N(\mu, \sigma^2)$ (a) shortest path and (b) nearest node

From the data, it can be observed that $g_{\eta=1}$ follows lognormal instead of normal distribution. To say that $\Delta X \sim N(\mu, \sigma^2)$ is clearly an inappropriate assumption. Recall that g is derived from ΔX. The results from $\Delta X \sim N(\mu, \sigma^2)$ are now put in question due to its assumption. High p indicates that majority of the vehicles are PV. While this could happen in the future, the dissertation is more interested at low p because this is the state of the current technology. Therefore it is important to analyze the results at low to mid p. After considering these two factors, it can now be observed that $\Delta X \sim N(\mu, \sigma^2)$ with nearest node produced the best results.

To predict η the shortest path will take the link with the minimum ε, which is the variance. As η increases the number of samples decreases affecting the variance. Instead of relying on the variance from limited number of samples, one approach is to assume linear relationship of the variance. Figure 60 illustrates this approach where a linear regression line is fitted through the

observations with high number of samples. The results shown previously are based on ε from linear regression.

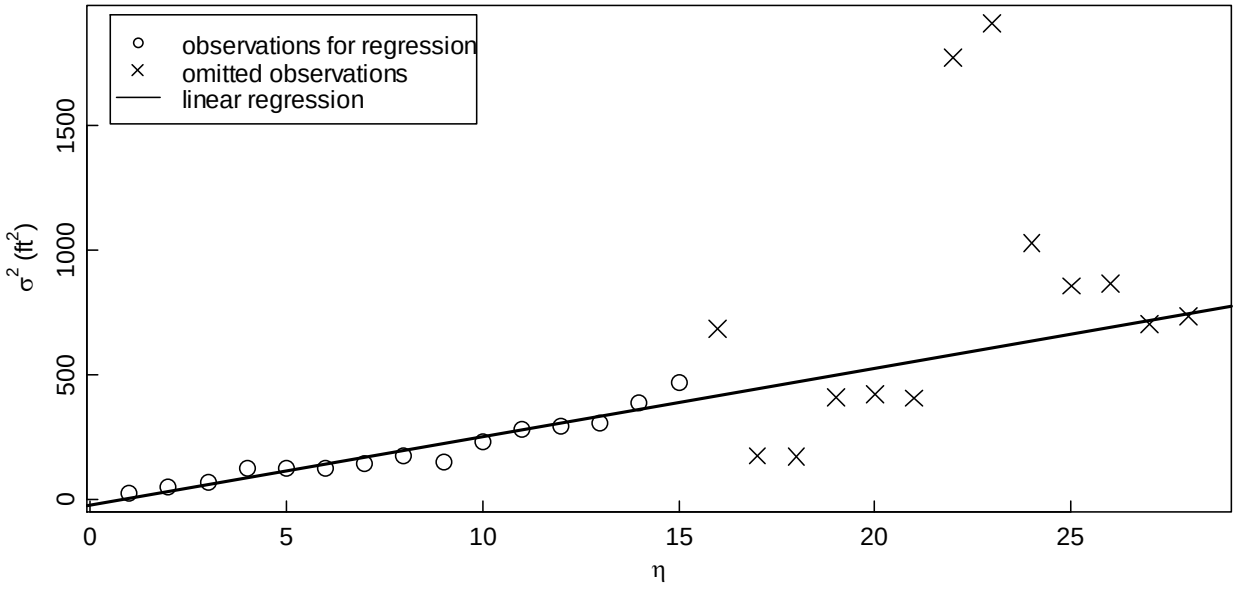

Figure 60 Variance and linear regression

Another option to assigning ε is the F-W approximation as shown in Figure 61. In this figure, the variance decreases as η increases. While this would result in different ε for each link, the F-W approximations requires that the distribution of g for all η to be lognormal. From previous analyses it cannot be said for certain that the distributions are lognormal. Similar to the linear assumption, the assumption of lognormal distributions is also excluded from analyses.

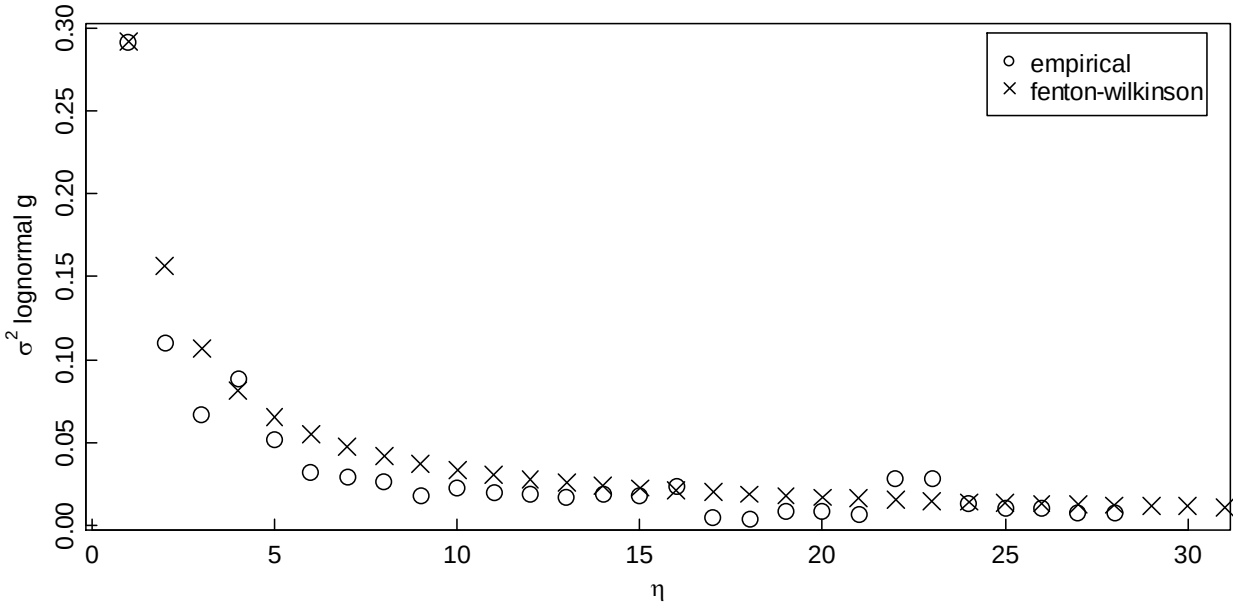

Figure 61 Fenton-Wilkinson approximation

As it stands, the results show that the accuracy for predicting flow is reasonable for the FD and SW approaches. While the PE of predicted η improves as p increase, the results from the CF approach can be improved. The list of possible improvements is being discussed in the next chapter. This concludes the explanation on the results of the proposed methodologies. The following chapter is an overall discussion of the dissertation.

CHAPTER 6

6 DISCUSSION

In this dissertation, three different methodologies are proposed for estimating traffic flow on freeways by combining PV data and different traffic flow theories namely fundamental diagram, shockwave and car-following. This chapter will summarize each of the methodology and points out their applicability and limitations.

For the FD approach, using the $q - u$ relationship, traffic volume can be estimated given a PV speed. For this method to work, a properly calibrated FD is required for the area of interest. PV speed at this area is then aggregated for a predefined time-space region. q is then estimated based on the PV speed.

q is estimated at different time intervals (5-, 10- and 15-minutes) for four different FDs each of which has its own unique $q - u$ relationship. Of the four FD, the Van Aerde model performed the best with the least prediction error. The reason is because the Van Aerde FD is fitted to the $q - u$ data instead of $u - k$ data as proposed by the other three FDs. Another finding is that as the aggregation interval increases from 5 to 15 minutes, the error decreases. By increasing the time interval, it's basically acting as a smoothing tool by reducing the fluctuation in the results. Finally, prediction error is low when traffic speed is high (free-flow) and error is high when speed is low (congestion). This is intuitive from observations of traffic flow where during free-flow traffic volume can fluctuate significantly while during congestion traffic flow has less fluctuation.

The FD approach is ready to be implemented by any transportation agency given the widespread availability of loop detector and PV data. One simply selects an area, fit a fundamental diagram from loop detector data and predict traffic volume from PV speed

However, the area of interest is limited to loop detector availability. Without loop detector data, a FD cannot be developed. If an area does not have a loop detector nearby, a FD from a similar roadway from a different area can be applied. With the assumption that the FD for both locations are similar. The prediction of traffic volume is dependent upon the fit of the FD. A properly calibrated FD is expected to give better prediction than a poorly fit FD or a one FD fits all approach.

While there have been studies on the PV-FD combination, the approach proposed in this dissertation is unique in terms of (1) Calibrated FD for the study area, (2) Investigation of four different models of FD and (3) Analyses for different aggregation intervals.

In the SW approach, traffic flow is divided into free-flow and congested regimes by performing *k*-means clustering to the PV data. SW is the ratio of difference in volume and difference in speed. From this relationship, there are five variables. Three of which (w, u_f and u_j) can be determined from PV data. The two remaining variables (q_f and q_j) are unknown. To solve the equation, q_j is estimated by using Northwestern congested regime FD. After performing this step q_f can now be calculated.

In this approach, with PV penetration rate of 5%, w is calculated for every 20th PV with the parameter u_B, which is introduced in the Northwestern congested regime FD. From the optimization process, $u_B$was found to be 35 mph. When plugged into the shockwave equation,

the results are unacceptable. As it turns out the Northwestern proposed value of $u_B = 40$ mph gave the best results.

This approach shows that traffic flow can be estimated when the traffic flow are divided into free-flow and congested regime. With the grouping of the traffic regime achieved by utilizing *k*-means clustering and predictions are made from the *w* formula.

There are limitations to this approach. First, without a shockwave or free-flow and congested regimes, this method is rendered useless. Different clustering technique will group the data differently affecting the *w* and the overall results. In this study *w* is calculated for every 20th PV. As penetration rate changes, this also affects the overall results. Too long of a period between PV (low penetration rate) might result in over smoothing the data while too short of period (high penetration rate) may show severe fluctuations. All of which affects the overall results.

The final methodology is the car-following approach which exploits the spacing between two vehicles identified as a leader and follower. In stop and go traffic flow, the state where a vehicle is stop is represented as flat line within its trajectory. When the stop motion of all vehicles are combined, they create a ladder shape within the trajectory. The distance between the horizontal lines inside the ladder are the vehicle spacing. During the training of the model, for each leader and follower pair, the distribution of vehicle spacing is determined. Using this distribution, predictions are made to the test data as to the number of vehicles between any two PV. During the testing, if intermediary PV exist between any two PV the prediction will take steps between the lead, intermediary and follower PV. The problem is solved in terms of finding the shortest path problem. The cost of the links that connect the vehicles are the variance of the spacing distribution as calculated from the training model.

From cross validation using NGSIM data it is shown that the error in predicting the follower number is reduced as the penetration rate increases. This is expected because the number of PV increases as penetration rate increases. With more PV the shortest path will have more links to choose from. The solution to the shortest path is dependent upon the link cost which is the variance of the gap. The nearest node method tends to overestimates the prediction of follower number due to the carryover of the estimated value from the lead to the follower.

In addition to the path selection method, another factor that contributes to prediction error is the transformation of spacing (front bumper to front bumper) to gap (back bumper to front bumper). Since vehicle length and the number of vehicles between two vehicles are not known, assumptions had to be made, which affects the predicted follower number.

When the follower is exactly behind the leader, the vehicle spacing has a lognormal distribution. However, the distribution becomes less definite if the follower is the second or third vehicle behind the leader. In fact the higher the follower position, the more difficult to identify its distribution. This happens due to variation in vehicle spacing and limited number of samples. For improve prediction, the distribution of the vehicle spacing needs to be determined properly.

In the end, the number of samples plays a crucial role in predicting the follower number. It affects the variance which used as the link cost and the distribution of vehicle gap. To improve prediction a higher number of samples are required.

The main limitations to this approach are that the lane assignment of the vehicles are known and lane changing do not occur. This information is typically not available for PV data. Additional steps are needed to determine the lane assignment from which one can determine if

lane changing occurs. In addition this approach only looks at passenger cars. Trucks are screened out prior to analyses.

For future studies, the shockwave and car-following approaches can be combined to estimate volume on a freeway. In the proposed shockwave approach, the volume in the congested region is estimated from the Northwestern fundamental diagram. For future study, this flow can be estimated from the car-following approach instead. Which will then be used to predict volume during free-flow by way of shockwave approach.

The car-following approach can be enhanced by investigating the proper distributions of the gap. An alternative to shortest path and nearest node can also be investigated.

CHAPTER 7

7 REFERENCES

Anderson, J., Ran, B., Jin, J., Qin, X., & Cheng, Y. (2011). Cycle-by-Cycle Queue Length Estimation for Signalized Intersections Using Sampled Trajectory Data. *Transportation Research Record: Journal of the Transportation Research Board, 2257*(-1), 87-94. doi:10.3141/2257-10

Antoniou, C., Ben-Akiva, M., & Koutsopoulos, H. (2004). Incorporating automated vehicle identification data into origin-destination estimation. *Transportation Research Record: Journal of the Transportation Research Board*(1882), 37-44.

Anuar, K., Habtemichael, F., & Cetin, M. (2015). *Estimating Traffic Flow Rate on Freeways from Probe Vehicle Data and Fundamental Diagram.* Paper presented at the Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on.

Ban, X., Hao, P., & Sun, Z. (2011). Real time queue length estimation for signalized intersections using travel times from mobile sensors. *Transportation Research Part C: Emerging Technologies, 19*(6), 1133-1156. doi:10.1016/j.trc.2011.01.002

Bar-Gera, H. (2007). Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. *Transportation Research Part C: Emerging Technologies, 15*(6), 380-391. doi:10.1016/j.trc.2007.06.003

Barceló, J., Montero, L., Marqués, L., & Carmona, C. (2010). Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring. *Transportation Research Record: Journal of the Transportation Research Board*(2175), 19-27.

Bartin, B., Ozbay, K., & Iyigun, C. (2007). Clustering-based methodology for determining optimal roadway configuration of detectors for travel time estimation. *Transportation Research Record: Journal of the Transportation Research Board*.

Beaulieu, N. C., & Rajwani, F. (2004). Highly accurate simple closed-form approximations to lognormal sum distributions and densities. *IEEE Communications Letters, 8*(12), 709-711.

Beaulieu, N. C., & Xie, Q. (2004). An optimal lognormal approximation to lognormal sum distributions. *IEEE Transactions on Vehicular Technology, 53*(2), 479-489.

Bertini, R. L., & Cassidy, M. J. (2002). Some observed queue discharge features at a freeway bottleneck downstream of a merge. *Transportation Research Part A: Policy and Practice, 36*(8), 683-697.

Blasch, E., Banas, C., Paul, M., Bussjager, B., & Seetharaman, G. (2012). *Pattern activity clustering and evaluation (PACE).* Paper presented at the SPIE Defense, Security, and Sensing.

Bucknell, C., & Herrera, J. C. (2014). A trade-off analysis between penetration rate and sampling frequency of mobile sensors in traffic state estimation. *Transportation Research Part C: Emerging Technologies, 46*, 132-150.

Caceres, N., Wideberg, J., & Benitez, F. (2007). Deriving origin destination data from a mobile phone network. *IET Intelligent Transport Systems, 1*(1), 15-26.

Cai, Q., Wang, Z., Zheng, L., Wu, B., & Wang, Y. (2014). Shock Wave Approach for Estimating Queue Length at Signalized Intersections by Fusing Data from Point and Mobile Sensors. *Transportation Research Record: Journal of the Transportation Research Board, 2422*(-1), 79-87. doi:10.3141/2422-09

Calabrese, F., Di Lorenzo, G., Liu, L., & Ratti, C. (2011). Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing, 10*(4), 0036-0044.

Cassidy, M., & Mauch, M. (2001). An observed traffic pattern in long freeway queues. *Transportation Research Part A: Policy and Practice, 35*(2), 143-156.

Cassidy, M. J., & Bertini, R. L. (1999). Some traffic features at freeway bottlenecks. *Transportation Research Part B: Methodological, 33*(1), 25-42.

Cetin, M. (2012). Estimating queue dynamics at signalized intersections from probe vehicle data. *Transportation Research Record: Journal of the Transportation Research Board, 2315*(-1), 164-172. doi:10.3141/2315-17

Cetin, M., & Nichols, A. (2009). Improving the accuracy of vehicle reidentification algorithms by solving the assignment problem. *Transportation Research Record: Journal of the Transportation Research Board*(2129), 1-8.

Chen, M., & Chien, S. (2000). Determining the number of probe vehicles for freeway travel time estimation by microscopic simulation. *Transportation Research Record: Journal of the Transportation Research Board*(1719), 61-68.

Chen, M., & Chien, S. (2001). Dynamic freeway travel-time prediction with probe vehicle data: Link based versus path based. *Transportation Research Record: Journal of the Transportation Research Board*(1768), 157-161.

Cobb, B. R., Rumi, R., & Salmerón, A. (2012). Approximating the distribution of a sum of log-normal random variables. *Statistics and Computing, 16*(3), 293-308.

Cohen, S., Bosseboeuf, J., & Schwab, N. (2002). *Probe vehicle sample sizes for travel time estimation on equipped motorways.* Paper presented at the Road Transport Information and Control, 2002. Eleventh International Conference on (Conf. Publ. No. 486).

Coifman, B. (2002). Estimating travel times and vehicle trajectories on freeways using dual loop detectors. *Transportation Research Part A: Policy and Practice, 36*(4), 351-364.

Comert, G., & Cetin, M. (2009). Queue length estimation from probe vehicle location and the impacts of sample size. *European Journal of Operational Research, 197*(1), 196-202. doi:10.1016/j.ejor.2008.06.024

Comert, G., & Cetin, M. (2011). Analytical evaluation of the error in queue length estimation at traffic signals from probe vehicledData. *IEEE Transactions on Intelligent Transportation Systems, 12*, 563-573.

Cullen, A. C., & Frey, H. C. (1999). *Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs*: Springer Science & Business Media.

Daganzo, C. F. (1994). The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research Part B: Methodological, 28*(4), 269-287.

Daganzo, C. F. (2005a). A variational formulation of kinematic waves: basic theory and complex boundary conditions. *Transportation Research Part B: Methodological, 39*(2), 187-196.

Daganzo, C. F. (2005b). A variational formulation of kinematic waves: Solution methods. *Transportation Research Part B: Methodological, 39*(10), 934-950.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik, 1*(1), 269-271.

Donovan, B., & Work, D. B. (2015). *Using coarse GPS data to quantify city-scale transportation system resilience to extreme events*. Paper presented at the Transportation Research Board, Washington, D.C.

Drake, J. S., Schofer, J. L., & May, A. D. (1967). *A statistical analysis of speed density hypotheses.* Paper presented at the Highway Research Record.

Edie, L. C. (1961). Car-following and steady-state theory for noncongested traffic. *Operations Research, 9*(1), 66-76.

Fenton, L. (1960). The sum of log-normal probability distributions in scatter transmission systems. *IRE Transactions on Communications Systems, 8*(1), 57-67.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The computer journal, 41*(8), 578-588.

Fujito, I., Margiotta, R., Huang, W., & Perez, W. (2006). Effect of sensor spacing on performance measure calculations. *Transportation Research Record: Journal of the Transportation Research Board*(1945), 1-11.

Greenberg, H. (1959). An Analysis of Traffic Flow. *Operations Research, 7*(1), 79-85. doi:doi:10.1287/opre.7.1.79

Greenshields, B. D. (1935). *A study in highway capacity.* Paper presented at the Highway research board.

Habtemichael, F. G., Cetin, M., & Anuar, K. A. (2015). Incident-Induced Delays on Freeways: Quantification Method by Grouping Similar Traffic Patterns. *Transportation Research Record: Journal of the Transportation Research Board*(2484), 60-69.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

Hao, P., Sun, Z., Ban, X. J., Guo, D., & Ji, Q. (2013). Vehicle index estimation for signalized intersections using sample travel times. *Transportation Research Part C: Emerging Technologies, 36*, 513-529.

Herrera, J. C., & Bayen, A. M. (2010). Incorporation of Lagrangian measurements in freeway traffic state estimation. *Transportation Research Part B: Methodological, 44*(4), 460-481.

Herrera, J. C., Work, D. B., Herring, R., Ban, X., Jacobson, Q., & Bayen, A. M. (2010). Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. *Transportation Research Part C: Emerging Technologies, 18*(4), 568-583. doi:10.1016/j.trc.2009.10.006

Hiribarren, G., & Herrera, J. C. (2014). Real time traffic states estimation on arterials based on trajectory data. *Transportation Research Part B: Methodological, 69*, 19-30. doi:10.1016/j.trb.2014.07.003

Hou, Y., Edara, P., & Sun, C. (2012). *A genetic fuzzy system for modeling mandatory lane changing.* Paper presented at the 2012 15th International IEEE Conference on Intelligent Transportation Systems.

Izadpanah, P., Hellinga, B., & Fu, L. (2009). *Automatic traffic shockwave identification using vehicles' trajectories.* Paper presented at the Proceedings of the 88th Annual Meeting of the Transportation Research Board (CD-ROM).

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters, 31*(8), 651-666.

Khodayari, A., Ghaffari, A., Kazemi, R., & Braunstingl, R. (2012). A modified car-following model based on a neural network model of the human driver effects. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 42*(6), 1440-1449.

Kianfar, J., & Edara, P. (2010). Optimizing freeway traffic sensor locations by clustering global-positioning-system-derived speed patterns. *IEEE Transactions on Intelligent Transportation Systems, 11*(3), 738-747.

Kim, S., & Coifman, B. (2014). Comparing INRIX speed data against concurrent loop detector stations over several months. *Transportation Research Part C: Emerging Technologies, 49*, 59-72. doi:10.1016/j.trc.2014.10.002

Kindzerske, M., & Ni, D. (2007). Composite nearest neighbor nonparametric regression to improve traffic prediction. *Transportation Research Record: Journal of the Transportation Research Board*(1993), 30-35.

Kurada, L., Öğüt, K., & Banks, J. (2007). Evaluation of N-curve methodology for analysis of complex bottlenecks. *Transportation Research Record: Journal of the Transportation Research Board*.

Kwong, K., Kavaler, R., Rajagopal, R., & Varaiya, P. (2009). Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors. *Transportation Research Part C: Emerging Technologies, 17*(6), 586-606.

Lam, C.-L., & Le-Ngoc, T. (2006). Estimation of typical sum of lognormal random variables using log shifted gamma approximation. *IEEE Communications Letters, 10*(4), 234-235.

Li, X., Jian, S., & Monteil, J. (2016). *Probe vehicle based technique to estimate fundamental diagrams on freeways and arterials*. Paper presented at the TRB 2016 Annual Meeting, Washington D.C.

Lighthill, M. J., & Whitham, G. B. (1955). *On kinematic waves II: A theory of traffic flow on long, crowded roads.* Paper presented at the Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences.

Lu, X.-Y., & Skabardonis, A. (2007). *Freeway traffic shockwave analysis: exploring the NGSIM trajectory data.* Paper presented at the 86th Annual Meeting of the Transportation Research Board, Washington, DC.

May, A. D. (1990). *Traffic flow fundamentals*.

Mehta, N. B., Wu, J., Molisch, A. F., & Zhang, J. (2007). Approximating a sum of random variables with a lognormal. *IEEE Transactions on Wireless Communications, 6*(7), 2690-2699.

Nanthawichit, C., Nakatsuji, T., & Suzuki, H. (2003). Application of probe-vehicle data for real-time traffic-state estimation and short-term travel-time prediction on a freeway. *Transportation Research Record: Journal of the Transportation Research Board*(1855), 49-59.

Navarro, M., & Herrera, J. (2014). Using Travel Time Data to Generate Aggregated Measures of Traffic. *Transportation Research Record: Journal of the Transportation Research Board*(2422), 96-103.

Neumann, T., Touko Tcheumadjeu, L. C., & Bohnke, P. L. (2013). *Dynamic representation of the fundamental diagram via Bayesian networks for estimating traffic flows from probe vehicle data.* Paper presented at the 16th International IEEE Conference on Intelligent Transportation Systems, Hague, Netherlands.

Neumann, T., Touko Tcheumadjeu, L. C., Bohnke, P. L., Brockfield, E., & Bei, X. (2013). *Deriving traffic volumes from probe vehicle data using a fundamental diagram approach*.

Paper presented at the 13th World Conference on Transport Research, Rio de Jeneiro, Brazil.

Newell, G. F. (1961). Nonlinear effects in the dynamics of car following. *Operations Research, 9*(2), 209-229.

Newell, G. F. (1993a). A simplified theory of kinematic waves in highway traffic, part I: General theory. *Transportation Research Part B: Methodological, 27*(4), 281-287.

Newell, G. F. (1993b). A simplified theory of kinematic waves in highway traffic, Part II: Queueing at freeway bottlenecks. *Transportation Research Part B: Methodological, 27*(4), 289-303.

Patire, A. D., Wright, M., Prodhomme, B., & Bayen, A. M. (2015). How much GPS data do we need? *Transportation Research Part C: Emerging Technologies, 58*, 325-342. doi:10.1016/j.trc.2015.02.011

Payne, H. J. (1971). Models of freeway traffic and control. *Mathematical models of public systems*.

Punzo, V., Borzacchiello, M. T., & Ciuffo, B. (2011). On the assessment of vehicle trajectory data accuracy and application to the Next Generation SIMulation (NGSIM) program data. *Transportation Research Part C: Emerging Technologies, 19*(6), 1243-1262.

Richards, P. I. (1956). Shock waves on the highway. *Operations Research, 4*, 42-51.

Romeo, M., Da Costa, V., & Bardou, F. (2003). Broad distribution effects in sums of lognormal random variables. *The European Physical Journal B-Condensed Matter and Complex Systems, 32*(4), 513-525.

Roncoli, C., Bekiaris-Liberis, N., & Papageorgiou, M. (2015). Highway traffic state estimation using speed measurements: case studies on NGSIM data and highway A20 in the Netherlands. *arXiv preprint arXiv:1509.06146*.

Schwartz, S. C., & Yeh, Y.-S. (1982). On the Distribution Function and Moments of Power Sums With Log-Normal Components. *Bell System Technical Journal, 61*(7), 1441-1462.

Seo, T., Kusakabe, T., & Asakura, Y. (2015a). Estimation of flow and density using probe vehicles with spacing measurement equipment. *Transportation Research Part C: Emerging Technologies, 53*, 134-150. doi:10.1016/j.trc.2015.01.033

Seo, T., Kusakabe, T., & Asakura, Y. (2015b). *Traffic state estimation with the advanced probe vehicles using data assimilation*. Paper presented at the IEEE 18th International Conference on Intelligent Transportation Systems, Gran Canaria, Spain.

Sethi, V., Bhandari, N., Koppelman, F. S., & Schofer, J. L. (1995). Arterial incident detection using fixed detector and probe vehicle data. *Transportation Research Part C: Emerging Technologies, 3*(2), 99-112.

Slimane, S. B. (2001). Bounds on the distribution of a sum of independent lognormal random variables. *IEEE Transactions on Communications, 49*(6), 975-978.

Srinivasan, K., & Jovanis, P. (1996). Determination of number of probe vehicles required for reliable travel time measurement in urban network. *Transportation Research Record: Journal of the Transportation Research Board*(1537), 15-22.

Sun, C., Ritchie, S. G., Tsai, K., & Jayakrishnan, R. (1999). Use of vehicle signature analysis and lexicographic optimization for vehicle reidentification on freeways. *Transportation Research Part C: Emerging Technologies, 7*(4), 167-185.

Sun, Z., & Ban, X. J. (2013). Vehicle trajectory reconstruction for signalized intersections using mobile traffic sensors. *Transportation Research Part C: Emerging Technologies, 36*, 268-283.

Treiber, M., Kesting, A., & Thiemann, C. (2008). *How much does traffic congestion increase fuel consumption and emissions? Applying a fuel consumption model to the NGSIM trajectory data.* Paper presented at the 87th Annual Meeting of the Transportation Research Board, Washington, DC.

Underwood, R. T. (1961). *Speed, volume, and density relationship: Quality and theory of traffic flow.* Paper presented at the Yale Bureau of Highway Traffic.

Single regime speed-flow-density relationship for congested and congested highways, (1995).

Van Aerde, M., & Rakha, H. (1995). *Multivariate calibration of single regime speed-flow-density relationships.* Paper presented at the Proceedings of the 6th 1995 Vehicle Navigation and Information Systems Conference.

Wang, H., Li, J., Chen, Q.-Y., & Ni, D. (2011). Logistic modeling of the equilibrium speed–density relationship. *Transportation Research Part A: Policy and Practice, 45*(6), 554-566.

Wei, D., & Liu, H. (2013). Analysis of asymmetric driving behavior using a self-learning approach. *Transportation Research Part B: Methodological, 47*, 1-14.

Work, D. B., Tossavainen, O.-P., Blandin, S., Bayen, A. M., Iwuchukwu, T., & Tracton, K. (2008). *An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices.* Paper presented at the Decision and Control, 2008. CDC 2008. 47th IEEE Conference on.

Zhao, L., & Ding, J. (2007). Least squares approximations to lognormal sum distributions. *IEEE Transactions on Vehicular Technology, 56*(2), 991-997.

VITA

Afi Anuar is a research assistant at the Old Dominion University Transportation Research Institute. He was admitted in 2009 earning a certificate in Maritime, Ports and Logistic Management (2010) and a Master's degree in Civil Engineering (2012). He is an active member of the Chi Epsilon and the Transportation Engineering Students Organizations. He held several positions as President and Secretary for both organizations.

Prior to joining ODU, he worked as an engineer in the automotive industry for nearly 10 years. In his free time, he enjoys photography and playing beach volleyball.