


2012

# Spatial and Temporal Correlations of Freeway Link Speeds: An Empirical Study

Piotr J. Rachtan

*University of Massachusetts Amherst*

Follow this and additional works at: <https://scholarworks.umass.edu/theses>

 Part of the [Applied Statistics Commons](#), [Civil Engineering Commons](#), [Engineering Science and Materials Commons](#), [Multivariate Analysis Commons](#), [Numerical Analysis and Computation Commons](#), [Other Operations Research](#), [Systems Engineering and Industrial Engineering Commons](#), [Statistical Models Commons](#), and the [Systems Engineering Commons](#)

---

Rachtan, Piotr J., "Spatial and Temporal Correlations of Freeway Link Speeds: An Empirical Study" (2012). *Masters Theses 1911 - February 2014*. 940.

Retrieved from <https://scholarworks.umass.edu/theses/940>

This thesis is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Masters Theses 1911 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact [scholarworks@library.umass.edu](mailto:scholarworks@library.umass.edu).

**SPATIAL AND TEMPORAL CORRELATIONS  
OF FREEWAY LINK SPEEDS: AN EMPIRICAL STUDY**

A Thesis Presented  
by

**PIOTR RACHTAN**

Submitted to the Graduate School of the  
University of Massachusetts Amherst in partial fulfillment  
of the requirements for the degree of

**MASTER OF SCIENCE IN CIVIL ENGINEERING**

September 2012

Civil and Environmental Engineering

**SPATIAL AND TEMPORAL CORRELATIONS  
OF FREEWAY LINK SPEEDS: AN EMPIRICAL STUDY**

A Thesis Presented

by

PIOTR RACHTAN

Approved as to style and content by:

---

Song Gao, Chair

---

Daiheng Ni, Member

---

Richard N. Palmer, Department Head

Civil and Environmental Engineering Department

## ACKNOWLEDGMENTS

The completion of this research would not be possible without the help of many people. First of all, I would like to express my gratitude to my research advisor, Professor Song Gao, who provided motivation, guidance and knowledgeable advice every time I needed it. Dr. Gao's doctoral student, He Huang, helped me at the early stages of this study by convincing me into programming and providing useful hints. I would also like to thank Professor Daiheng Ni for serving at the thesis committee, but also for his questions and opinions on the matters of his specialty.

I would like to extend my thanks to Professor John Staudenmayer for his willingness to consult statistical matters; and to Professor John Buonaccorsi for his continuous engagement into helping me to resolve programming problems in SAS and R. In special cases, I could always count on Tonya Chapman from SAS Technical Support, whose replies were always very quick and answered my follow-up questions before I knew I would need to ask them.

Another thank you should go to California Department of Transportation which operates PeMS. This study would not be possible at any stage without the data they have provided.

I ought to thank my parents, Anna and Roman, who tirelessly fostered my return to school. A special thanks to my wife, Karolina, for her love, support and encouragement – I would not be where I am and who I am without her. Finally, I would like to thank my daughters, Amelia and Helena, for their patience when I could not play and their smiles whenever I needed them the most.

## ABSTRACT

### SPATIAL AND TEMPORAL CORRELATIONS OF FREEWAY LINK SPEEDS: AN EMPIRICAL STUDY

SEPTEMBER 2012

PIOTR RACHTAN

A.S., SPRINGFIELD TECHNICAL COMMUNITY COLLEGE  
B.S.C.E., UNIVERSITY OF MASSACHUSETTS AMHERST  
M.S.C.E., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Song Gao

Congestion on roadways and high level of uncertainty of traffic conditions are major considerations for trip planning. The purpose of this research is to investigate the characteristics and patterns of spatial and temporal correlations and also to detect other variables that affect correlation in a freeway setting. 5-minute speed aggregates from the Performance Measurement System (PeMS) database are obtained for two directions of an urban freeway – I-10 between Santa Monica and Los Angeles, California. Observations are for all non-holiday weekdays between January 1st and June 30th, 2010. Other variables include traffic flow, ramp locations, number of lanes and the level of congestion at each detector station. A weighted least squares multilinear regression model is fitted to the data; the dependent variable is Fisher Z transform of correlation coefficient.

Estimated coefficients of the general regression model indicate that increasing spatial and temporal distances reduces correlations. The positive parameters of spatial and temporal distance interaction term show that the reduction rate diminishes with spatial or temporal distance. Higher congestion tends to retain higher expected value of correlation; corrections to the model due to variations in road geometry tend to be minor. The general model provides a framework for

building a family of more responsive and better-fitting models for a 6.5 mile segment of the freeway during three times of day: morning, midday, and afternoon.

Each model is cross-validated on two locations: the opposite direction of the freeway, and a different location on the direction used for estimation. Cross-validation results show that models are able to retain 75% or more of their original predictive capability on independent samples. Incorporation of predictor variables that describe road geometry and traffic conditions into the model works beneficially in capturing a significant portion of variance of the response. The developed regression models are thus transferrable and are apt to predict correlation on other freeway locations.

# TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS.....	iii
ABSTRACT.....	iv
LIST OF TABLES.....	viii
LIST OF FIGURES.....	ix
CHAPTER	
1. INTRODUCTION.....	1
1.1 Background.....	1
1.2 Literature Review.....	1
1.3 Research Objectives.....	4
1.4 Expected Contribution.....	5
1.5 Paper Organization.....	5
2. PRELIMINARY ANALYSIS (PART I): I-10 EASTBOUND.....	6
2.1 Setting and Data Description.....	6
2.2 Spatial Correlation Analysis.....	7
2.2.1. Methodology.....	7
2.2.2 Results.....	8
2.3 Temporal Correlation Analysis.....	10
2.3.1 Methodology.....	10
2.3.2 Results.....	11
2.4 Adaptive Link State Method.....	14
2.5 Preliminary Regression.....	15
2.5.1 Model.....	15
2.5.2 Evaluation.....	17
3. PRELIMINARY ANALYSIS (PART II): I-880 NORTHBOUND.....	19

3.1 Setting and Data Description.....	19
3.2 Spatial Correlation Analysis.....	19
3.3 Temporal Correlation Analysis.....	20
4. PRELIMINARY FINDINGS SUMMARY.....	22
5. ANALYSIS – FINAL STAGE.....	23
5.1 Background and Motivations.....	23
5.2 Methodology and Literature Review on Correlation Modeling.....	24
5.3 Data.....	26
5.3.1 Final Analysis Setting.....	26
5.3.2 Filtering.....	26
5.3.3 Additional Information.....	27
5.4 Regression.....	30
5.4.1 Pool of Predictor Variables.....	30
5.4.2 Model Selection Procedure.....	31
5.4.2.1 24-Hour, Full Segment Models.....	31
5.4.2.2 Daytime, Full Segment Models.....	32
5.4.2.3 Full-Day, Partial Segment Model.....	33
5.4.2.4 Three Models, Partial Segment.....	33
5.4.3 Model Diagnostics and Results.....	34
5.4.4 Morning Model Evaluation.....	37
5.4.5 Midday Model Evaluation .....	39
5.4.6 Afternoon Model Evaluation.....	41
6. CROSS-VALIDATION.....	44
6.1 Background and Motivations.....	44
6.2 Location 1: I-10 Westbound.....	44
6.3 Location 2: I-10 Eastbound (Beginning).....	47
6.4 Cross-Validation Summary.....	50
7. OTHER MODELING STRATEGIES.....	52
8. CONCLUSIONS AND RECOMMENDATIONS.....	55
8.1 Summary.....	55
8.2 Recommendations for Future Research.....	55



APPENDICES.....57

1. LAYOUT SCHEMATICS OF INTERSTATE 10 BETWEEN PM 0 AND PM 12.50,  
BOTH DIRECTIONS.....58

2. DATA FILTERING PROCESS – FREQUENCY OF % OBSERVED DATA.....59

3. SPEED-FLOW DIAGRAMS FOR I-10 E-W WITH CONGESTION  
THRESHOLDS INDICATED.....62

    A.EASTBOUND DIRECTION.....62

    B. WESTBOUND DIRECTION.....69

4. TRAFFIC CONDITION FREQUENCY TABLES FOR DAYTIME PERIODS.....74

BIBLIOGRAPHY.....81

## LIST OF TABLES

Table	Page
1. Preliminary Regression Results for Three Cases.....	16
2. Stable/Unstable Condition Thresholds for I-10 E Stations .....	29
3. Stable/Unstable Condition Thresholds for I-10 W Stations .....	29
4. Summary of the Final Models Estimated for Partial Segment of I-10 E .....	35
5. Predictive Ability of the Regression Models on Three Independent Locations .....	50
6. Performance Indicators of Station-to-Station Models .....	53

## LIST OF FIGURES

Figure	Page
1. Spatial Correlation Patterns for Link 1 with All Other Links under Varied Link Definitions.....	9
2. Spatial Correlations for Link 1 with All Other Links Calculated Using the Base Setup; a) During Peak, b) During Off-Peak Periods.....	10
3. Temporal Correlation Patterns a) During, and b) After Morning Rush Hour.....	11
4. Temporal Correlations Into- and After the Peak.....	12
5. Temporal Correlation Patterns for 7:00, 8:30, and 10:00 with All Other Periods Over the Entire 6-Hour Analysis.....	13
6. Spatial Correlation Patterns for a) Link 1, and b) Link 6 with All Other Links on I-880 N.....	20
7. Temporal Correlation Patterns for a) 2:00 pm and b) 4:00 pm with All Other Time Periods.....	21
8. Morning Model: Observed vs. Predicted Spatial Correlation. a) PM=7.05 at 9:00, b) PM=10.79 at 9:00.....	38
9. Morning Model: Observed vs. Predicted Temporal Correlation. a) PM=6.86 at 10:00, b) PM=8.07 at 7:00.....	39
10. Midday Model: Observed vs. Predicted Spatial Correlation. a) PM=5.64 at 12:00, b) PM=10.79 at 12:00.....	40
11. Midday Model: Observed vs. Predicted Temporal Correlation. a) PM=5.64 at 14:00, b) PM=8.07 at 14:00.....	41
12. Afternoon Model: Observed vs. Predicted Spatial Correlation. a) PM=7.05 at 17:00, b) PM=8.38 at 17:00.....	42
13. Afternoon Model: Observed vs. Predicted Temporal Correlation. a) PM=6.86 at 15:00, b) PM=10.07 at 17:00.....	43
14. Spatial Correlation Cross-Validation, Observed vs. Predicted (I-10 W). a) Morning: PM=6.58, 9:00, b) Midday: PM=10.45, 12:00, c) Afternoon: PM=11.96, 17:00.....	45

15. Temporal Correlation Cross-Validation, Observed vs. Predicted (I-10 W). a) Morning: PM=11.96, 8:30, b) Midday: PM=10.79, 13:00, c) Afternoon: PM=9.38,15:00.....46

16. Spatial Correlation Cross-Validation (I-10 E, beginning). Observed vs. Predicted. a) Morning PM=2.35, 9:00, b) Midday: PM=4.00, 12:00, c) Afternoon: PM=4.00, 17:00.....48

17. Temporal Correlation Cross-Validation (I-10 E, beginning). Observed vs. Predicted. a) Morning: PM 0.17, 10:00; b) Midday: PM=0.17, 14:00; c) Afternoon: PM=1.77, 16:00.....49

# CHAPTER 1

## INTRODUCTION

### 1.1 Background

According to Transportation Research Board (2009), congestion is a growing problem on the urban highways over the world, and it is becoming worse with the increasing number of commuters and random disruptions to the system, such as incidents, road work, bad weather, etc. In order to model traveler's route choice decisions and provide reliable prediction of future traffic conditions along the chosen path, a stochastic time-dependent network is required to capture the uncertainties.

There usually exist strong stochastic dependencies among link speeds (or travel times), largely due to traffic flow propagations over time and space, or an event that affects capacities in a wide area. Network stochastic dependencies are generally required to capture the benefits of real-time information for network routing, since only through the dependencies over time and space can the knowledge of an incident at the current time result in a better prediction of traffic conditions in the future at different location within the network. However, the shape of correlation patterns is still largely unknown and, as the literature review shows, most of the research in the related areas either base on simplifying assumptions or ignore link correlation.

### 1.2 Literature Review

This section aims to present an overview of recent research efforts that are related to our link correlation study. Prediction of short-term future traffic condition on real-time basis can allow travelers to avoid existing or likely congestion. Zhang and Rice (2003) develop a linear prediction model with time-varying coefficients, but do not factor correlations into the procedure. Rice and van Zwet (2004) continue work on travel time prediction and use historical

travel times to estimate regression parameters. They present a method to predict travel times that is computationally effective, but estimated level of congestion is based on an assumption that travel times for a given travel path aim to their historical mean. By accounting for correlation, the prediction could be perhaps more computationally intensive, but more adaptive to traffic conditions changing during the trip. In their recent work, Samaranayake, Blandin and Bayen (2011) use a Bayesian network framework to learn its time-space dependencies using a structure learning algorithm. The algorithm is a simplified version of greedy-equivalence search algorithm based on assumptions that simplify traffic state evolution. Algorithm is tested using freeway loop-detector data to test the short-term forecasting accuracy of speed and travel times.

Others, like Gajewski and Rilett (2004) take correlation into account. They estimate link travel time correlation using Bayesian natural cubic splines – a nonparametric regression technique. From all mentioned, this is the only empirical study that tries to quantify the correlations that are not assumed, but from actual traffic data. Note that the correlations studied here are at the link level with aggregate traffic and day-to-day randomness, rather than at the vehicle level, as studied in Gajewski and Rilett with probe-vehicle data. Chandra and Al-Deek (2008) investigate the effect of upstream and downstream location information by checking cross-correlation of speeds at these locations relative to the current location of hypothetical traveler. They find significant relationship of speed with stations both upstream and downstream of the traveler. Tam and Lam (2009) use historical travel time estimates together with their updated temporal variance-covariance relationships to predict the travel times in the next five-minute interval, and they show that use of the updated temporal variance-covariance relationships of travel times can greatly improve the accuracy of the short-term travel time prediction. Min and Wynter (2010) develop volume and speed forecasting models using

binomial spatio-temporal correlations based on simplified link condition - traffic condition may be either congested or free flowing. Although based on simple assumptions, their model achieves reasonable accuracy in short-term prediction of speed and volume.

Adaptive routing, investigated – among others – by Waller and Ziliaskopoulos (2002), who approach the problem with limited forms of spatial and temporal link cost dependencies. Given the cost of predecessor links, no further information is obtained through spatial dependence; limited temporal dependency assumes known link cost when the entrance node is reached. In contrast, Gao and Chabini (2006) study optimal routing policy problems with an assumption of complete dependencies; they recognize that capturing link correlations over time and space would potentially make the route choice models more realistic.

Other important research areas that deal with link correlation are studies on volume and Origin-Destination (OD) demand forecasting. Goel et al. (2005) explore correlations between 24-hour segment volumes and prove that including correlation improves Average Annual Daily Traffic (AADT) prediction; however, large ratio of OD pairs to links may result in overestimation of correlation coefficients. Eom et al. (2006) develop a spatial regression model which considers spatial dependency effect (correlation). Their regression model gains overall predictive capability and accuracy over the ordinary least squares regression when strong correlations exist. Song et al.(2009) model correlation between OD demands during given time period (e.g. morning peak). They show that mean traffic flows are very sensitive to correlation changes and thus correlations significantly influence travelers' path choice behaviors. A method to evaluate uncertainty of future demand is shown by Duthie et al. (2011).

Transportation planning and policy are also sensitive to spatial and temporal dependencies within a transportation network. Frejinger and Bierlaire (2007) capture correlation

among alternatives in a route choice problem. They introduce subnetwork to simplify the road network based on the original network's road hierarchy. The modified multinomial logit model they propose captures correlation among paths by error components. Parent and LeSage (2010) develop a space-time dynamic model that relates commuting times with highway infrastructure, gasoline taxes and congestion. They find that spillover of spatial effects is substantially stronger than time impacts. They also find that neglecting of positive or negative correlations in analysis of variance of future travel times may lead to underestimation as large as 75%, or its overestimation by 100%, respectively.

### **1.3 Research Objectives**

This research is a study of link speed correlations in a freeway setting. It aims to document the following:

- to quantify correlations and measure the dependencies with regard to time and space separately using loop detector data from freeway corridors, determine how sensitive calculations are to link definition and time resolution;
- to investigate correlation patterns over time and space simultaneously through a regression model;
- to assess factors other than time and space that affect correlation;
- to discuss the transferability of the findings to other freeway locations through the means of cross-validation;
- to assess the potential of extending the research to arterial roadways or ideally a road network.



## **1.4 Expected Contribution**

The contribution of this study is to fill the gap of lack of empirical quantification of spatio-temporal link correlation patterns. This study sets an important first step in explaining link correlation phenomena and may potentially serve as a springboard to develop more realistic models in areas such as: travel time prediction, origin-destination demand forecasting, adaptive routing, and transportation policy evaluation and planning.

## **1.5 Paper Organization**

The somewhat unusual structure of this thesis reflects the development and evolution of the research behind it. The ideas and motivations for improvement were grounded in the preliminary results (Chapters 2, 3 and 4). The actual final analysis is described and evaluated in Chapters 5 and 6. Chapter 7 overviews an exploratory approach to modeling that uses samples only from adjacent stations. This is followed with conclusions, future research directions and contribution in the last chapter.

## CHAPTER 2

### PRELIMINARY ANALYSIS (PART I): I-10 EASTBOUND

#### 2.1 Setting and Data Description

The data for the primary model are obtained from California Department of Transportation Performance Measurement System (PeMS). The location chosen for development of the initial study is a 12.04 mile (19.38 km) segment of I-10 E freeway in Los Angeles vicinity. It stretches from mile post 0.17 in Santa Monica to mile post 12.21 at the intersection with I-110 in Los Angeles. The primary criterion for choosing this location is heavy congestion occurring on a daily basis with high predictability and along a considerable stretch of highway.

We choose to analyze only non-holiday weekdays from 7 am to 1 pm when the congestion is severe and occurs with high regularity. 5-minute speed data aggregates from period between March 1 2010 and June 30 2010 from 7:00 am to 12:59:59 pm are obtained for a total of 87 weekdays from all 41 loop detector stations operated by PeMS along this segment.

Initial scoping of the data sets detects unrealistic speed readings on stations that have been out-of-order during the study period. Those data were produced by the “imputation” algorithm implemented by PeMS. We decide to ignore those stations’ readings and filter out data from other stations (some of which have been shut down for several days); thus, only high quality readings (detector health at 50% or better) are used for numerical analyses. Filtering process of data reduces the number of acceptable detectors to 34.

## 2.2 Spatial Correlation Analysis

### 2.2.1 Methodology

As commonly known and studied from the earliest years by traffic engineering researchers, congested traffic behavior differs greatly from traffic behavior in uncongested state. Peak hour for this I-10 E segment has been determined by finding a significant drop in speed that occurs with some regularity between 7:30 and 9:30 am. Separate analyses are performed for peak hour (7:30-9:30 am) and off-peak hours (9:30 am-12:59 pm). For this stage of research, peak/off-peak period have been specified by visual inspection of speed/flow plots on randomly selected links and days. Mean peak hour speed  $x_p=39.38$  mph (63.40 km/h), mean off-peak speed  $x_{OP}=56.49$  mph (90.95 km/h). In our future work, we may consider using flow information to set a threshold above which link is considered congested.

5-minute speed aggregates are first grouped using one of the link setups presented below; mean values (over peak, or off-peak hour) of aggregates are taken across each link generating a single mean speed value at each day for each link. This is followed by the removal of remaining missing values that survived the filtering process (usually less than 5%). Then, Pearson's correlation coefficients between each pair of the average link speeds are calculated, resulting in a  $n$  by  $n$  correlation matrix, where  $n$  is the number of links. The sample size in calculating correlation coefficients is the number of days (87) less missing values (explained below).

Given the filtered data, several link definitions are considered to explore the influence of data grouping on the correlations. First two setups are grounded in the physical features of the freeway and the latter two are based on the more arbitrary use of detector stations. The following link definitions are investigated:

- *Base ramp-to-ramp setup* – a link is defined as a stretch of the road between detector stations in the closest proximity of an exit ramp; all intermediate stations are averaged in the link speed. The boundary detector readings are used for link ( $i$ ) as well as for link ( $i+1$ ). The segment divides into 13 links when this definition is used.

- *Variation 1.* Alternative ramp-to-ramp setup has the same basis as the base model with an exception of overlap – links start at a ramp and end before next ramp. Thus, boundary detector readings are used just once. Since the number of ramps does not vary, there are 13 links generated by this setup. Both variation 1 and base setup are referred to as “ramp-to-ramp setups” later in text.

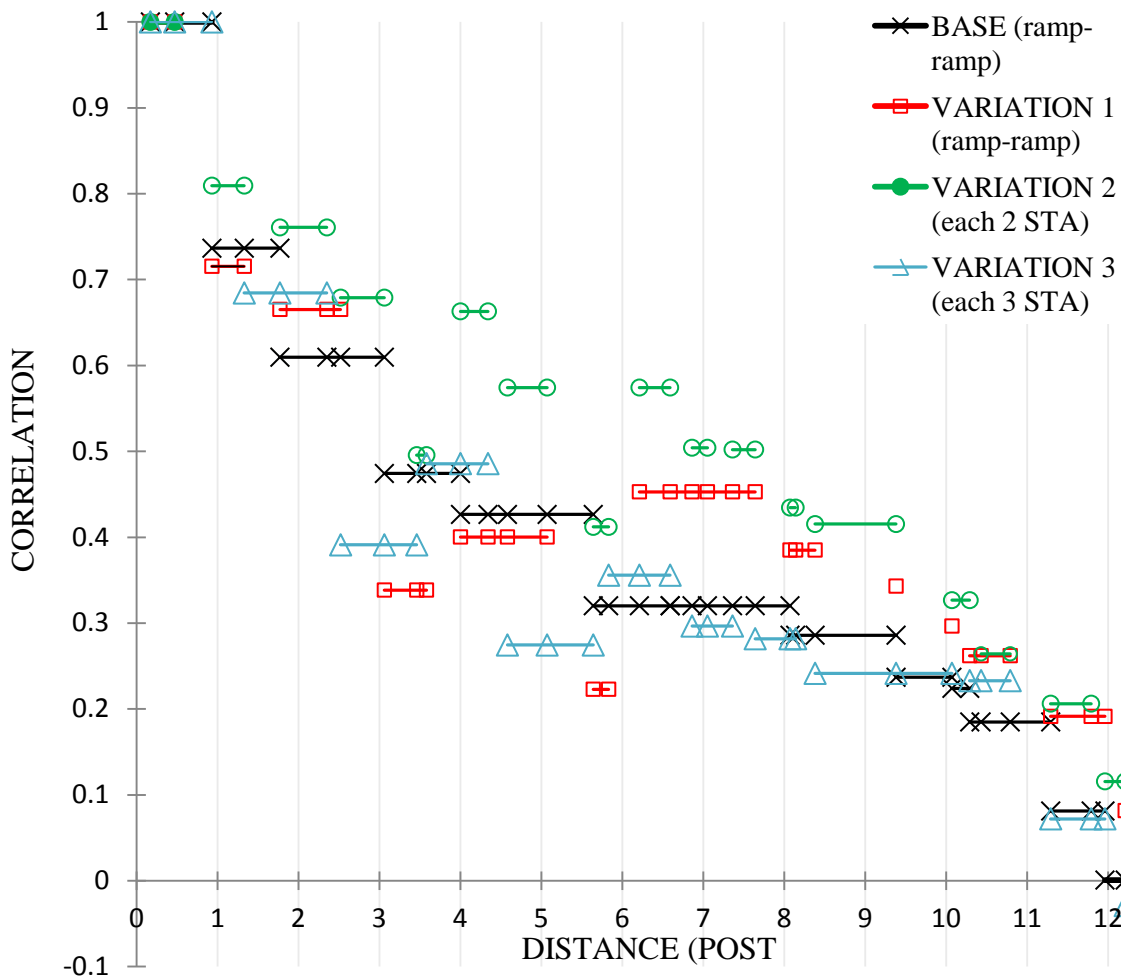
- *Variation 2.* A link is defined as a stretch of the road between two consecutive detector stations regardless of the distance. An average link speed is always of two detector data. This method yields 17 links.

- *Variation 3.* Each link ‘contains’ 3 detector stations (with the exception of the last link which is only 1 station). 12 links generated by this definition share similarities with variation 2 (arbitrary number of stations used), but also with ramp-to-ramp setups since link length is often similar (approximately 1 mile).

## **2.2.2 Results**

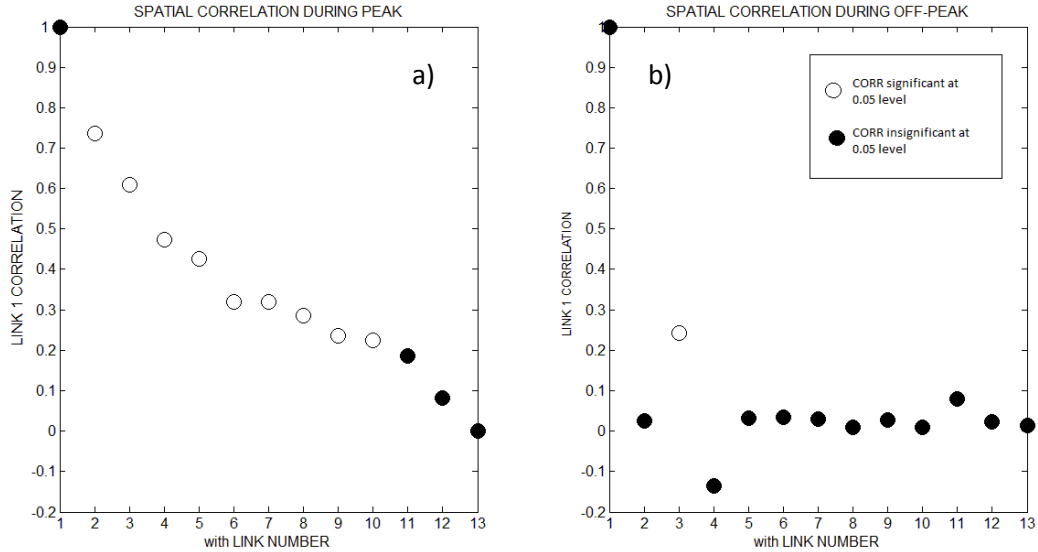
Following the investigation of the influence of varied link definitions on correlation pattern, Figure 1 depicts spatial correlation results for link 1 and other links using all four setups. Note that horizontal axis reflects distance in miles rather than link numbers. Horizontal lines represent links as defined by each setup; markers are detector stations.

Ramp-to-ramp setups (base and variation 1) are preferred over the other two more arbitrary models because they follow freeway facilities and are more intuitive. In some cases, though, ‘arbitrary’ models may be preferred because their setup on the large scale can possibly be automated. Regardless of the setup used, the same trend is reflected in link correlations – link correlations ultimately drop to zero over the course of 12.04 miles. However, some differences exist: link correlation coefficients calculated with either ramp-to-ramp model (base and variation 1) tend to fit in the middle between the other two setups. Base setup seems to produce a smoother correlation pattern; therefore it is used for the analyses that follow.



**Figure 1** Spatial Correlation Patterns for Link 1 with All Other Links under Varied Link Definitions

The calculated spatial correlation coefficients for link 1 with all other links for base ramp-to-ramp setup are gathered in Figure 2 with regard to peak and after peak situations.



**Figure 2 Spatial Correlations for Link 1 with All Other Links Calculated Using the Base Setup; a) During Peak, b) During Off-Peak Periods**

In both cases link correlations obviously drop with increased link distance; however, it seems that strong correlations are farther reaching during peak as off-peak correlations very quickly start to oscillate around zero. In addition, all peak correlations above 0.2 in value are statistically significant at the 0.05 level – that is not the case for off-peak. For peak period, sample size  $N=84$  observations (days); for off-peak period  $N=80$  (depending on the amount of missing values that needed to be removed from the initial 87 days).

## 2.3 Temporal Correlation Analysis

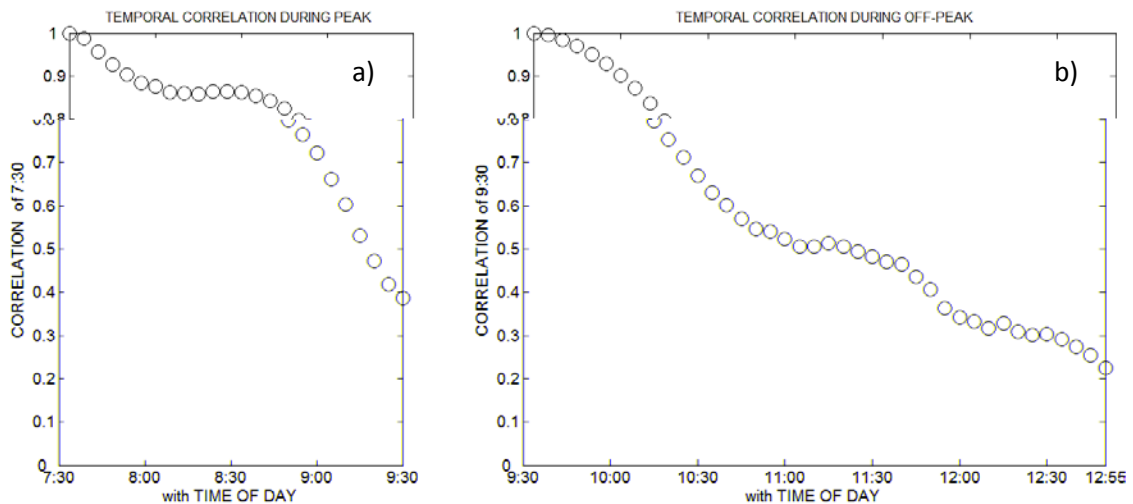
### 2.3.1 Methodology

To minimize spatial influence on correlation, link definitions are ignored and a mean speed value across entire segment is taken. Similarly to the spatial analysis which deals with a few setups, four distinctive time periods are used in temporal analysis: 60-minute, 30-minute, 15-minute, and

raw 5-minute aggregates are used for calculating mean segment speed at different resolution. Following the methodology used for spatial analysis, any missing data is removed at a cost of slight reduction of sample size (number of days) to 79 from original 87 observations. Size of the square correlation matrix depends on the resolution and its dimension varies  $\{[6 \times 6], [12 \times 12], [24 \times 24], \text{ or } [72 \times 72]\}$ , for 60-minute, 30-minute, 15-minute and 5-minute resolution, respectively.

### 2.3.2 Results

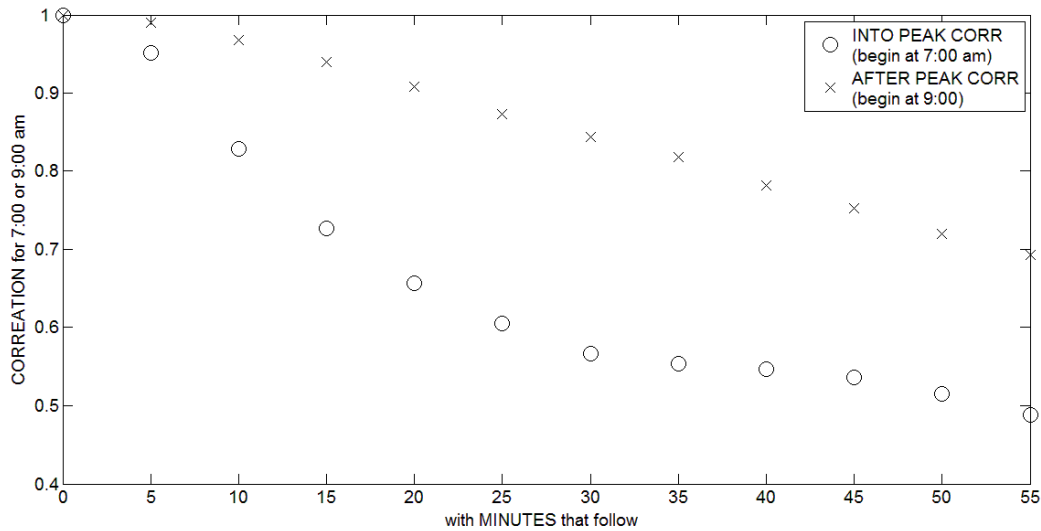
Similarly to the spatial correlation analysis, temporal correlation is studied for peak hour and the off-peak periods to investigate various correlation patterns under different traffic conditions. Temporal correlation pattern does not seem to significantly vary with the choice of resolution. Figure 3 shows temporal correlation patterns during 2-hour period during the morning rush and 3.5-hour period after the morning rush with respect to the beginning of each period. These plots are at 5-minute resolution.



**Figure 3** Temporal Correlation Patterns a) During, and b) After Morning Rush Hour

In both cases correlations are strongest within the first 30 minutes, which follows general intuition. Off-peak correlations tend to drop after that initial period, while peak correlations remain at high level ( $>0.8$ ) for about 70 minutes deep into the peak (7:30-8:40) before a steep drop after that. Still, 2 hours during and 3.5 hours after morning rush seem to be too little time to result in no correlation. With sample sizes  $N=84$  for peak period and  $N=80$  after the morning peak, all correlation coefficients tend to be statistically significant at the 0.05 level.

It may be interesting to investigate how does correlation change during the free-flow-congestion and congestion-free-flow transitions of traffic condition. Figure 4 depicts temporal correlations patterns for 7:00 am and 5-minute intervals within an hour (until 8:00 am) – when congestion is believed to build-up – plotted against correlations for 9:00 am and corresponding intervals within an hour (until 10:00 am), when congestion is assumed to dissipate. Horizontal axis represents 5-minute intervals that begin as marked and end 4:59 (min:sec) later (e.g. “0” means period from 7:00 to 7:04:59; “55” means period from 7:55 to 7:55:59; and similarly for periods starting at 9:00 am).

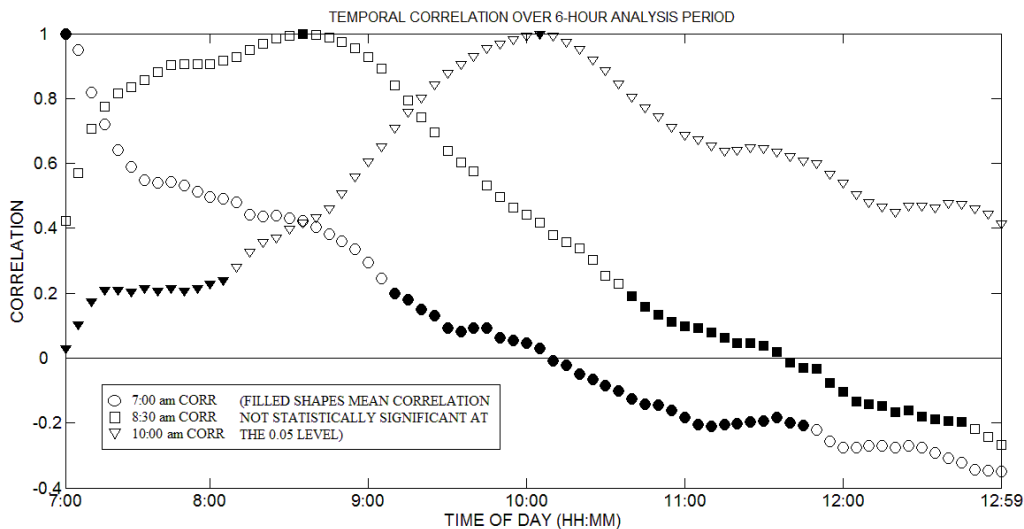


**Figure 4** Temporal Correlations Into- and After the Peak



For sample size  $N=86$  for into peak to  $N=85$  for after peak period, correlation coefficients are all statistically significant at the 0.05 level. It seems that speeds are much higher correlated when congestion dissipates than while it builds up. Near linear correlation drop is observed during after peak period and are still strong (near 0.7) after 1 hour. In contrast, correlation pattern during congestion build-up drops steeper and forms a sag-like shape. Within 30 minutes (this period coincides with 7:30 am, which is the beginning of peak and assumed congestion), correlation drops near the value of 0.5 when the pattern begins to flatten out. Possible explanation of the difference in correlation patterns at the edges of congestion is that congestion dissipation is a steadier process – there is less speed variability and dynamic changes in traffic condition than during congestion build-up.

Above analyses lack an answer to an important question: when does correlation reach zero? To observe this temporal correlation drop, it is necessary to use the whole 6 hour analysis period for calculations. Figure 5 depicts temporal correlations for analysis period between 7:00 am and 12:59:59 pm.



**Figure 5** Temporal Correlation Patterns for 7:00, 8:30, and 10:00 with All Other Periods Over the Entire 6-Hour Analysis

As expected, correlation coefficients plot for 7:00 am, 8:30 am, and 10:00 am reveals that – over long enough time – temporal correlation not only drops to zero but also reaches negative values. Negative correlation occurs between 7:00 am (before rush hour) and 12:55 pm (after morning rush) with a value -0.351; similarly, the correlation between 8:30 am (rush) and 12:55 pm takes a value of -0.2685. Both results are statistically significant for sample size  $N=72$ , despite the loss of significance when correlations are near-zero, roughly in the correlation range  $\{-0.2, 0.2\}$ . We suspect this unexpected finding may be due to fluctuations in flow, i.e. when more commuters choose to travel during or before peak hour (thus reducing mean speed in the early hours), less commuters do travel after the morning rush (thus speed is higher than average off-peak speed) as the total demand on one direction tends to remain constant. This hypothesis will be investigated in future work as it requires all-day data, preferably for both directions. It is also possible that this phenomenon is location specific to I-10 E, as temporal correlation drop is not observed in our investigation of I-880 N presented in Chapter 3.

Another interesting phenomenon in Figure 5 is the different slopes of the 8:30 am correlation curve before and after 8:30. To the left, there is a slow decrease and then a sharp decrease, while to the right there is a steady and moderate decrease; though 90 minutes away from 8:30 correlation on both sides reaches the same level of 0.4. No negative correlation occurs on the 10:00 am (after peak) curve. The slope before 10:00 am is much steeper (i.e. correlation drops more rapidly) than after 10:00 am. Some of the findings will be further discussed following the regression model in latter parts of Section 2.5.

## **2.4 Adaptive Link State Method**

Somewhat arbitrary determination of whether links are congested or not – as described in section 2.1 – is convenient, but offers very crude accuracy in terms of the real traffic condition. After

obtaining flow data, the actual traffic conditions on each link may be determined by setting a threshold quantity that will distinguish whether a link is congested or uncongested at any given time based on historical data for a link. This method is expected to provide day-specific adjustments to make correlation calculations adaptive to actual traffic conditions. In general, this approach follows an idea described in Jia et al. (2000), who suggested setting a threshold on loop detector occupancy below which free flow is assumed.

## **2.5 Preliminary Regression**

### **2.5.1 Model**

Another approach to data organization is taken for development of the regression model for correlation prediction. The regression is done for time-dependent link speeds and combines the two separate, 1-dimensional analyses into a 2-dimensional (space and time) unified model. There is a mean speed value for each link at each time stamp (total of  $13 \times 72 = 936$ ) per each of the 87 observations (days). Base setup (described in section 3.2.1) is used as link definition. Missing data has not been removed but rather interpolated from adjacent readings from same detector stations. This procedure, also implemented by Rice and van Zwet(2004), is preferred at this stage since only a few missing values would mean unnecessary and troublesome removal of many thousand readings.

Based on this data set, correlation coefficients form a 936 by 936 element matrix. Since the matrix is symmetric, only elements from the lower triangle (including the diagonal of ones) are used for estimation of regression coefficients. Sample size  $N=438516$ , which is yielded by  $[(936)^2/2 + 936/2]$ , is valid for the responses (correlation coefficients) and is also the dimension of prediction parameters. A multilinear linear regression is fitted to the data and its parameter estimates are presented in Table 1.

**Table 1 Preliminary Regression Results for Three Cases**

N=438516		R <sup>2</sup> =0.5687							
constant (fixed)	distance	time_diff	distance* time_diff	distance* OP_dum my	time_diff* OP_dum my	distance* time_diff* OP_dummy	distance* OO_dum my	time_diff* OO_dum my	distance* time_diff* OO_dummy
1	-0.0634	-0.00330	+0.000312						
				-0.0107	-0.000168	+0.0000647			
							+0.00626	+0.000754	-0.0001933
t-test	-328.01	-489.39	210.32	-36.861	32.766	-20.595	105.97	29.387	-68.336
standard error	1.93E-04	6.74E-06	1.484E-06	2.90E-04	0.0001911	8.192E-06	7.103E-06	2.201E-06	1.746E-06

The model is applicable to 3 distinctive cases:

- Peak–peak, which is the base case and works when correlation coefficient is calculated between two time intervals, both of which are during peak, or morning rush hour. The first row of parameters is applied in this case;

-Peak–off-peak, applicable when calculating correlation coefficient between time interval during peak and time interval before/after the peak, or opposite. The second row of parameters provides corrections to the three parameters in the base case;

-Off-peak–off-peak, this case covers all situations that relate pre-peak and after peak time intervals, including relations within the same interval (pre-peak to pre-peak). This case requires using the third row of parameters as corrections to the base case from Table 1.

Regardless of which case is considered, the first constant has been fixed to 1 to reflect the fact that correlation with self (for zero temporal and spatial distances) is equal to 1 by definition.

The following variables are used in the regression model:

- *distance* – number of links between the requested link and the link with which correlation is calculated {0,1,2....12}; “0” is for correlation within a single link, “12” is for correlation between first and last link.

- *time\_diff* – time difference in minutes between requested 5-minute interval and the interval with which correlation is calculated {0,5,10,15...355}; these values represent time difference to a beginning of correlated interval. Again, “0” is for correlation within a single 5-minute interval, while “355” is between intervals of 7:00-7:04:59 with 12:55-12:59:59.
- *OP\_dummy* – a dummy variable with values of “0” for peak-peak and off-peak–off-peak situations, or “1” when a combination of peak–off-peak situations occurs.
- *OO\_dummy* – a dummy variable with values of “0” for peak-peak and peak–off-peak situations, or “1” when a combination of off-peak–off-peak situations occurs.
- the remaining variables are interaction terms that are simple multiplication of parameters in other variables; as indicated in Table 1.

### **2.5.2 Evaluation**

All cases reflect correlation drop with increase of distance and time difference. Peak-peak case is controlled primarily by spatial distance, as the temporal term is of much smaller value. Off-peak–peak situation tends to inflate temporal influence by almost 54%; yet, spatial term inflated by 17% is still much more significant in combined absolute value. Much steeper shape of temporal correlations during the first 60 minutes in Figure 5 (Off-peak–peak situations) when compared to left plot in Figure 3 (Peak-peak situations) agrees with this regression result. Off-peak–off-peak case shows weakened influence of increasing difference in time (23% reduction) and space (10% reduction). Weaker drop of correlation over time is visible when two plots in Figure 3 are compared. Positive sign of the interaction term of distance and time difference indicates that it slows down the decrease of correlation with increase of spatial or temporal distance. This result confirms intuition, as correlation is caused primarily by flow propagation over finite time. As a result, the correlation between link 1 at 8:00 am and link 2 at 8:05 am is

potentially stronger than either that between links 1 and 2 at 8:00 am or that between link 1 at 8:00 am and 8:05 am, since the vehicles on link 1 at 8:00 am and those on link 2 at 8:05 am are probably more or less the same.

There are two possible explanations of the larger influence of spatial distance. Along the 12.04 mile (19.38 km) stretch, freeway geometry varies (changing number of lanes, different on- and off-ramp designs and so on), so different locations (links) can dissolve correlations quickly. Second, large on- and off-ramp traffic is likely to disturb the mainline traffic and make links less related across distance. Imagine an extreme case when there are no access ramps along a segment of considerable length, and the spatial correlation in such situation should be higher. On the other hand, temporal correlation is simpler and might only depend on the spatial distribution of demand.

## CHAPTER 3

### PRELIMINARY ANALYSIS (PART II): I-880 NORTHBOUND

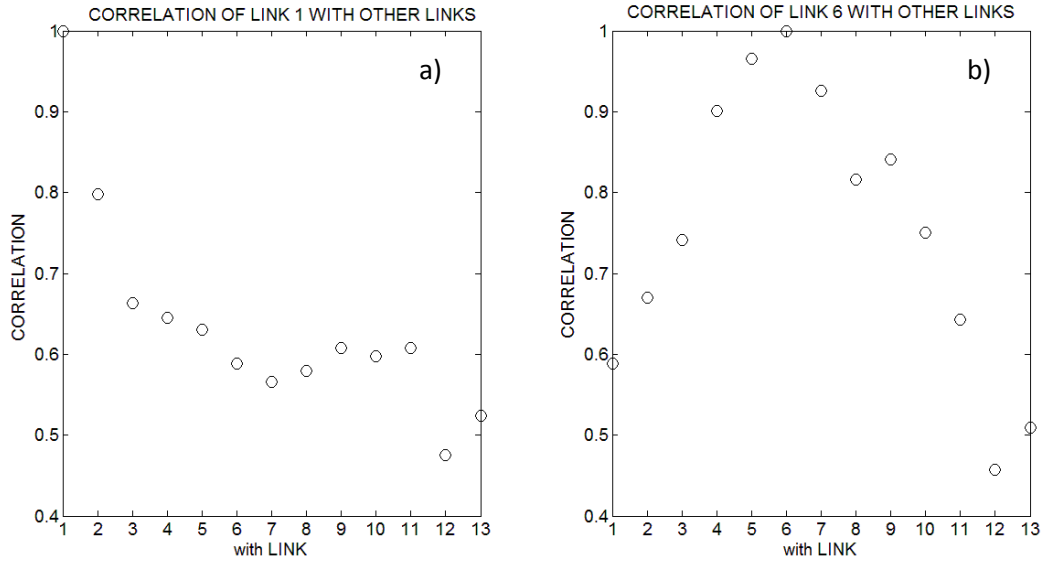
#### 3.1 Setting and Data Description

To investigate the transferability of the initial findings to another freeway, a second analysis is done on the I-880 N (Nimitz Freeway) segment between Fremont (exit 15) and San Leandro (exit 31). The segment is 16.16 miles (26.02 km) long, stretching between detector stations at mile posts 14.89 and 31.05. PeMS data is gathered for the same 87 weekdays between March 1 2010 and June 30 2010.

Inspection of speed patterns on random days in the chosen period indicated hardly an existence of a morning rush hour on this segment; instead, a long afternoon rush is observed. Therefore, 7 hour period from 1:00 pm to 8:59:59 pm is selected. Since a drop (or, more accurately – a strong variability) of aggregated speeds seems to occur roughly over the aforementioned 7 hour, we assume that entire period as afternoon rush and decide to omit off-peak cases. There are 35 detectors along this stretch. Data quality is much better and requires significantly less data filtering.

#### 3.2 Spatial Correlation Analysis

Following the previously defined “base setup”, links stretch from ramp to ramp. This method generates 13 links on I-880 N. Data is prepared as in section 3.2.1., with the difference of 7-hour period (instead of 6 hours for I-10 E). Spatial correlation is calculated after removal of missing values, and sample size  $N=73$ . Figure 6 shows spatial correlation coefficients for links 1 and 6 on left and right plot, respectively.



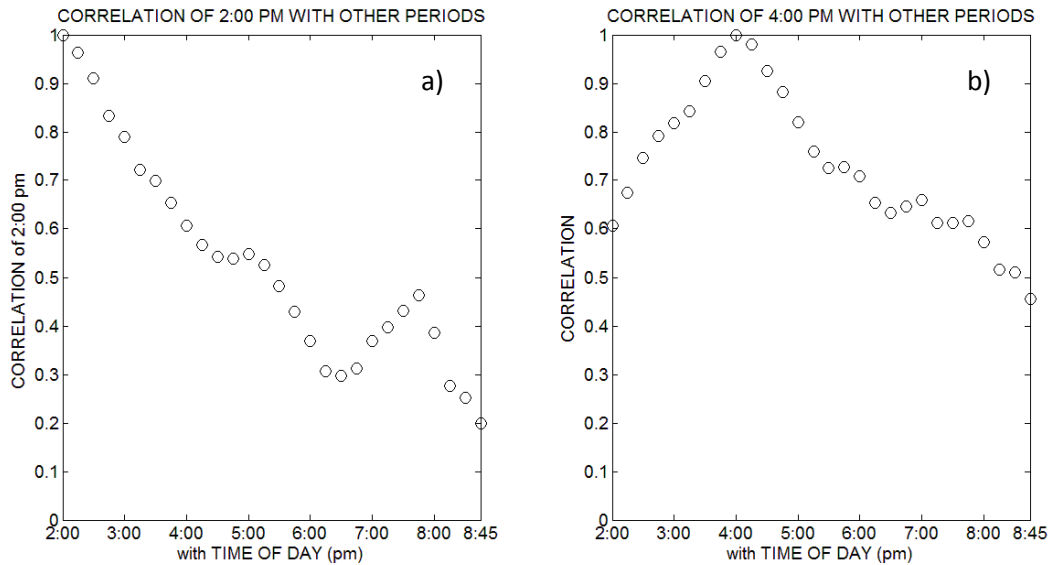
**Figure 6 Spatial Correlation Patterns for a) Link 1, and b) Link 6 with All Other Links on I-880 N**

Correlation seems to drop to about 0.5 – 0.6 level with the difference of just about 2 links. This is true for both plots, as well as the fact that no correlation values below roughly 0.45 appear. A quick comparison to Figure 2 reveals that spatial correlation patterns for I-880 N are significantly different from that of I-10 E. Even if the shape itself resembles that of off-peak correlation plot in figure 2, values are significantly higher and all correlations in Figure 6 are statistically significant at the 0.05 level.

### 3.3 Temporal Correlation Analysis

The methodology for calculating temporal correlations for I-10 E described earlier is followed for the I-880 N data set. Different resolutions in time dimension, however, have not been studied and only 15-minute resolution is available. As in previous section, no off-peak period has been differentiated within the 7 hour data set. Temporal correlation patterns are shown in figure 7.





**Figure 7 Temporal Correlation Patterns for a) 2:00 pm and b) 4:00 pm with All Other Time Periods**

Note that all correlations are statistically significant at the 0.05 level for both plots. This is a major difference with discussion following Figure 5. Another one is that temporal correlation drops more abruptly on I-10, while I-880 does not reach near-zero values; not to mention the lack of observed existence of negative correlation. The correlation patterns in Figure 7 are more resembling of those in Figure 3, and correlations reach similar values. An important note needs to be mentioned, however, that plots in Figure 3 are done for 2 and 3.5 hour periods – not for a 7 hour period as those in Figure 7; the difference in scale is very important. Given the considerable differences in correlation patterns between the two freeways, we believe that link correlation is likely to be location specific.

## CHAPTER 4

### PRELIMINARY FINDINGS SUMMARY

In this preliminary research, traffic data from two urban freeways are obtained using the PeMS database and analyzed to study the characteristics of stochastic dependencies among link speeds. All analyses confirmed that correlation of link speeds drops over temporal and spatial distances. We have shown that correlation patterns are not very affected by link definition choice for analysis; however, it may be viable to drop the link definition whatsoever and try to work on raw data instead of the averages.

Separating spatial and temporal dimensions gives an insight into each aspect while minimizing the effect of the other. Performing multivariate linear regression on combined spatio-temporal data confirms the hypothesis that spatial distance has stronger negative effect on correlation than temporal distance, regardless of the overall segment traffic condition. Separate spatial and temporal analyses have been reiterated using the data from another location and their results suggest that link correlation patterns are location specific.

Although we have strived to follow our informed intuition to do the described work most accurately, we have found many venues for improvement. The flaws detected in the preliminary methodologies are listed together with motivations for improvement in the opening section of the following chapter.

## CHAPTER 5

### ANALYSIS – FINAL STAGE

#### 5.1 Background and Motivations

The exploratory character of the preliminary analysis allowed for some simplifications that now should be corrected in the development of a valid model and valid conclusions. First, the regression model was estimated for bounded and clustered response (correlation), which violated the assumption of constant error term variances (Neter et al., 1996).

Secondly, the initial study revealed the need to increase sample size. On many occasions, it was difficult to tell if a large p-value on correlation coefficient estimates was a result of insufficient sample size or simply indicated correlation that indeed was not different from zero. Increasing the sample size is aimed to address these doubts.

Thirdly, the preliminary analysis is done based on a specific definition of link; several such definitions have been explored, but all share a tendency to stabilize the correlation variation by using certain averaging scheme. Although the approach of using several stations' readings as single link reading tends to improve fit of any regression model, such fit inflation may be sensitive to the link definition applied by the modeler. Moreover, the ramp-to-ramp link definition used to estimate the preliminary regression model has a different number of detector stations in each link. Thus, prediction error may vary with actual spatial distance. Again, the error variance is not constant for different levels of the predictor variable. For the above reasons, it was decided to use raw detector station readings instead of links.

In addition, as mentioned in previous sections, one of the key parameters that intuitively affects correlation – level of congestion – however based on exploration of historical speeds at several detector stations, it has been determined somewhat arbitrarily. Given many detector

stations along the analyzed segment and day-to-day or even minute-to-minute variation in traffic conditions, it is believed that such variation should be reflected in the predictor variables at an attainable level. This should improve the fit of the model by making it more responsive to traffic condition changes regardless if the given time of day is peak or off-peak.

Lastly, preliminary regression model does not take road geometry into account. The existence of on- or off-ramps has been known in traffic engineering to create disturbances to the traffic stream. On the other hand, increased number of mainline lanes reduces the density and vehicle-to-vehicle interaction. Therefore, inclusion of such information to the model may result in model's increased robustness when it is transferred to a new location with different characteristics. On top of all other considerations, we believe this study should result in a model (or models) that can be valid for the entire day instead of just a peak hour.

## **5.2 Methodology and Literature Review on Correlation Modeling**

In order to develop a regression model that satisfies the theoretical requirements, the methodology needs to be supported by research in the field of linear modeling of correlation coefficient.

A transformation given by Fisher (1928) is required to obtain a response variable from correlation that will ensure the homogeneity of variance, will be unbounded and its distribution will be approximately normal:

$$w(z) = \frac{1}{2} \ln \left[ \frac{1+r(z)}{1-r(z)} \right], \text{ where } w(z) \text{ is the Fisher Z transform and } r(z) = r_{yx}(z) \text{ is}$$

sample correlation coefficient estimated variable pair  $y$  and  $x$  for the given  $z$ .

According to Bartlett (1993), for moderately large samples we may assume that  $w(z)$  is approximately normally distributed and that the variance of  $z$ :

$Var(w(z)) \cong \frac{1}{n-3}$ , where  $n$  denotes sample size for correlation coefficient

estimation.

Bartlett points out; however, that Fisher Z transformation alone does not eliminate unequal variances of the error terms even though it is normally distributed, as confirmed by simulations described in Asuero, Sayago and González (2006). Since the variance depends on the sample size behind each sample correlation estimation, application of weighted least squares regression with weights  $W_k = (n_k - 3)$ , where  $k=1,2,\dots,K$  and  $K$  is the total number of pairs.

To compare observed correlation to the predicted values produced by a regression model, transformation of Fisher Z's back to correlation domain is necessary. To perform back-transformation, the following formula is applied:

$$\hat{r}(z) = \frac{\exp(2E(z))-1}{\exp(2E(z))+1} \quad , \text{ where } \hat{r}(z) \text{ is the correlation predicted from } E(z), \text{ the expected}$$

value of  $z$ .

To assess the general fit of a regression model to the observed sample correlation values (in this analysis: response  $y_i$ ), three statistical measures can be applied as follows:

- Coefficient of determination  $R^2 = \frac{SSM}{SST} = \frac{SSM}{SSE + SSM} = \frac{\sum_i(\hat{y}_i - \bar{y})^2}{\sum_i(y_i - \hat{y}_i)^2 + \sum_i(\hat{y}_i - \bar{y})^2}$ , where

$SST$  is the total sum of squares,  $SSM$  is the sum of squares in the model and  $SSE$  is the sum of squares error;  $\hat{y}_i$  is prediction of the  $i$ -th realization of the response  $y_i$ , and  $\bar{y}$  is the mean of the response.

- Bias as an estimate of accuracy,  $bias = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)$ , where  $n$  denotes the number of observations of the response.

- Estimate of precision,  $Standard\ error = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$ .

## **5.3 Data**

### **5.3.1 Final Analysis Setting**

Similarly to the initial study, data in the actual analysis are 5-minute speed aggregates obtained by loop detectors operated by PeMS. Same 12.04 mile (19.38 km) stretch is used as in the initial analysis setting, but the eastbound segment of I-10 between PM (post mile) 0.17 and PM 12.21 (a total of 41 detector stations) has been supplemented by the westbound segment of I-10 between PM 11.96 and PM 0.17 (a total of 34 detector stations). Including the westbound direction serves several purposes: it assists the model building process by providing additional ability to consult another data set; it also provides an opportunity to transfer the model onto a location that has a different geometrical arrangement and experiences different traffic conditions. The geometrical arrangement of this bi-directional segment is shown in Appendix 1.

As mentioned in Section 5.1, initial analysis revealed a need to increase the sample size. Now, after expanding the study period by another two months, it stretches from January 4<sup>th</sup>, 2010 to June 30<sup>th</sup>, 2010; a total of 127 non-holiday weekdays within this 6-month period. In addition, data are now for 24 hours, not just the morning rush hour. More observation days increase the sample size but also reduce the susceptibility for reporting interruptions. 24-hour data allows for estimation of more general model. Although all detectors experience short maintenance shutdowns over the relatively long six month period, some detectors have not been operating for extended periods of time. Such situation causes another potential variance stabilizing interference, as PeMS still does report the interpolated readings from problem stations.

### **5.3.2 Filtering**

Filtering process begins with an assessment of data quality at each station. Quality reported by PeMS describes the percentage of detectors reporting for a given station; e.g. if a reading for a 3-

lane freeway is 67%, it means that during that time period only 2 out of 3 detectors are working. First step of the filtering process involved the initial screening of the data for each station: if there are no observed data (i.e. quality=0%) for the entire period, then such station is removed from the data set. In this step, 7 stations for I-10 E and 4 stations for I-10 W were removed from further analysis. As indicated by the diagnostics of the filtering process (Appendix 2), a large number of observations with 0% report quality remained after the first filtering step. Following with step two, missing values are assigned in place of reported speed when data quality is less than 50%. Such action does not interfere with the time stamp for each observation, but missing values do not go into further analyses. To avoid losing large amounts of observed data, 50% data quality is chosen as a reasonable estimate of the traffic conditions. For a 4-lane highway, if one knows the situation on just two lanes, one can assume other two lanes are not very different; in other words, situations when one can encounter congested and free-flowing traffic on different lanes of the same direction at the same location are very rare.

### **5.3.3 Additional Information**

Manual input is needed to include information not directly reported in the raw data. First of all – since the link definition is dropped and speed readings are associated with specific locations – it is necessary to include physical location in post miles, not just use station ID's. Using reports for both directions (PeMS Station Info I-10 E-W, 2012), post mile corresponding to each station ID for all stations that passed the first step of filtering process is assigned as variable *post\_mile*.

From the same source tables, new variable *ramp* is created. *Ramp* value is equal to 1 if a station is associated with a presence of a ramp (no distinction is given between access and egress), *ramp* is equal to 0 otherwise.

Again, using the same station inventory reports, variable *num\_lanes* is also incorporated into the data sets. Its value is equal to the number of mainline lanes at a given station.

As first mentioned in the preliminary analysis report, the congestion level is often location specific; thus, new variable *state* is introduced through a multi-step procedure. It begins with taking both speed and total flow (aggregated from all lanes) readings to develop speed-flow diagrams for each station using best-possible data quality, preferably and in most cases above 75%. Speed-flow plots, which are enclosed as Appendix 3, served to determine speed thresholds at which congestion occurs at a given station. In general, the attempt was to choose a speed reading which can conservatively indicate unstable traffic conditions (speed smaller than speed at capacity; in some cases taking care not to include lower speeds at the free-flow branch of a plot). The thresholds for I-10 E stations are listed in Table 2; thresholds for I-10 W stations are listed in Table 3.

With the stable/unstable condition thresholds assigned to each station using data from the entire analysis period, within each 5-minute reporting period determined is a proportion of unstable periods. This number is a continuous variable *state*, bounded between 0 (indicating 100% stable conditions) and 1 (100% unstable conditions). The variable satisfies the requirements of an ‘adaptive’, localized indicator of traffic condition.



**Table 2      Stable/Unstable Condition Thresholds for I-10 E Stations**

<b>Speed Thresholds I-10 E</b>			
Post Mile	Speed	Post Mile	Speed
0.17	72	6.59	54
0.47	66	6.86	60
0.93	61	7.05	71
1.33	58	7.36	62
1.77	60	7.64	62
2.35	57	8.07	53
2.52	61	8.14	56
3.06	60	8.38	65
3.46	63	9.38	61
3.59	58	10.07	62
4.00	58	10.29	61
4.34	60	10.43	45
4.58	63	10.79	63
5.07	64	11.29	62
5.64	50	11.79	70
5.83	63	11.96	64
6.21	53	12.21	62

**Table 3      Stable/Unstable Condition Thresholds for I-10 W Stations**

<b>Speed Thresholds I-10 W</b>			
Post Mile	Speed	Post Mile	Speed
11.96	62	6.21	60
11.79	75	5.66	60
11.51	62	5.07	60
10.79	58	4.58	60
10.45	55	4.30	62
10.07	62	4.00	60
9.38	68	3.46	60
9.04	62	3.06	52
8.90	62	1.97	62
8.38	61	1.77	60
8.14	56	0.93	62
7.64	61	0.78	59
7.05	60	0.47	61
6.58	60	0.17	74

## 5.4 Regression

### 5.4.1 Pool of Predictor Variables

From parameters introduced in the previous section, several derivative variables can be created after noting that estimation of each correlation coefficient pertains to a pair of speed readings from two time-space locations over a given sample size (number of observations available for that pair). Variables used in the model building are as follows:

- *Timediff* – similar to preliminary definition (*time\_diff*); for sample correlation  $r_{yx}$ , *timediff* is the temporal distance in minutes between  $y$  and  $x$ . Due to the resolution of raw data, *timediff* increment is 5-minutes.
- *Miles* – in contrast to the preliminary study, where *distance* indicated number of links between locations  $y$  and  $x$ . *Miles* is the absolute value of the difference in post miles for locations  $y$  and  $x$ , one whole unit of spatial distance is equal to 1 mile (1.61 km).
- *State* – for the purpose of regression, variable *state* is the mean value of state at location  $y$  and state at location  $x$ .
- *OP\_dummy* – ‘peak-off-peak’ indicator variable derived from *state* in such a manner that traffic at a station is assumed congested when  $state \geq 0.5$ , uncongested otherwise. Then, *OP\_dummy* is equal to 1 if one, and only one of  $y$  and  $x$  locations is congested; otherwise, *OP\_dummy* is equal to 0.
- *OO\_dummy* – ‘off-peak-off-peak’ indicator variable based on the same simplification of stable/unstable traffic condition determination. *OO\_dummy* is equal to 1 if both  $y$  and  $x$  are uncongested; otherwise, *OO\_dummy* is equal to 0.
- *lanes* – road geometry parameter, follows the same logic as *state*; it is a mean *num\_lanes* from locations  $y$  and  $x$ .

- *RN\_dummy* – another road geometry parameter; this ‘ramp-no-ramp’ indicator variable is based on the values of *ramp* in locations *y* and *x*. *RN\_dummy* is equal to 1 if one, and only one of *y* and *x* locations is adjacent to ramp; otherwise, *RN\_dummy* is equal to 0.
- *NN\_dummy* – this ‘no-ramp-no-ramp’ indicator variable complementary to *RN\_dummy* in the same manner as *OO\_dummy* complements *OP\_dummy*. *NN\_dummy* is equal to 1 if none of *y* and *x* locations is adjacent to a ramp; otherwise, *NN\_dummy* is equal to 0.

Please note that *state* is mutually exclusive with *OP\_dummy* and *OO\_dummy*, as they convey the same information.

## 5.4.2 Model Selection Procedure

### 5.4.2.1 24-hour, Full Segment Models

Various combinations of variables are used to estimate a model for 24-hour data on the entire I-10 E segment. The primary objective is to achieve best fit using variable combinations that make physical sense; the secondary objective is to obtain the simplest model possible. To agree with the postulates laid out in Section 5.2, all models are using weighted least squares method; the response variable is Fisher Z transform.

Model building process starts with an additive model that only includes *timediff* and *miles*; then the basic model is expanded by interaction term, and then by squares of *miles* and *timediff*. Bartlett (1993) recommends using the quadratic model when variability in sample correlation is roughly uniform and estimated on continuous variable. These apply in this study and the model with both interaction term and quadratic terms has the best fit from all models using just spatial and temporal distance as predictors. These models are later referred to as ‘simple’. The best simple model achieved adjusted  $R^2=0.221$ .

Next family of models, later referred to as ‘advanced’ incorporates usage of the additional variables from the pool to the simple model with the best fit. Advanced models are built by sequential addition of variables and interaction terms. For models to correctly behave when spatial and temporal distances are 0 (self-correlation) by leaving just the non-zero intercept, additional variables have to be interacted with either spatial or temporal distance; i.e. cannot exist ‘on their own’ in the model. Multiple scenarios are developed, including second degree interaction terms and interaction of additional variables with quadratic terms. The biggest model boasted 36 covariates; however, it did not offer significantly better fit than much smaller and more transparent models. The best general advanced model, with its adjusted  $R^2=0.254$  achieved slightly better fit than the best simple model.

#### 5.4.2.2 Daytime, Full Segment Models

The general models’ significantly worse performance to the model in the preliminary analysis is believed to have many reasons, among which:

- the models try to describe relationships in the raw data, not the averages, as there are no links and observations are station-specific and vary more;
- there are no ‘imputed’, interpolated values in the data;
- night-time traffic is believed to have flows and traffic densities incomparably smaller than during daytime; also it is primarily free-flowing with correlations not significantly different from zero.

Since the first two causes are also among the primary motivations for the final analysis, they cannot be avoided. However, we decided that including night-time information in the model serves little purpose, and, by over-generalization on low-flow data regression loses its ability to

capture fine-scale variability of Z transform. Data is thus scaled to the 14-hour period between 6:00 and 20:00.

Iterative model fitting similar to that of described in the previous section resulted in the significant improvement in fit for both simple model (adj.  $R^2=0.358$ ) and advanced model (adj.  $R^2=0.361$ ). Careful examination of the advanced model's prediction of temporal and spatial correlation on observed vs. predicted (where predicted correlation is back-transformed from expected values of Fisher Z's) plots revealed that the day-time model is still too general to capture sudden changes in correlation.

#### 5.4.2.3 Full Day, Partial Segment Model

The segment's physical arrangement during the first five post miles (later referred to as 'beginning') seems to be quite different from what follows between Venice Boulevard (PM=5.64) and I-110 Interchange (PM=11.96). At the latter fragment, it is difficult to find any two consecutive stations without any ramps. It also seems to be congested more often than the beginning. Therefore, it was decided to use the latter fraction of the segment as the base for further model estimation in hope of improving the models' prediction performance.

Downscaling the spatial dimension of the sample resulted in further improvement of the fit of the advanced model, which attained  $R^2=0.465$ . Although the fit at this stage is far better than that of the original model described in Section 5.4.2.1, we noticed a pattern of improving model performance with each reduction of estimation data. The sample sizes are still very large and thus significant reductions do not lead to noticeable weakening of parameter estimates.

#### 5.4.2.4 Three Models, Partial Segment

We have decided to follow with an investigation of how the traffic conditions look during each 5-minute period over the 6-months of analysis. For this purpose, contingency tables are prepared,

in which percentage of stable/unstable conditions at all stations (on I-10 E, full segment) are shown for each 5-minute period of day (e.g. period number 73 indicates 6:00-6:04:59; period number 85 indicates 7:00-7:04:59, etc.). These tables serve the purpose of finding the beginning and end of typical rush hours during the day, but also to investigate the traffic condition changes within rush hours. It was found that peak hours are characterized by prolonged congestion, morning rush hour starts at 7:30 and ends at 9:50; afternoon peak starts at 13:25 and ends at 19:45. Peak was determined by finding periods with percentage of congestion greater or equal to 50%. These tables are enclosed as Appendix 4.

Since each cut to the data resulted in a model that is more adaptive to variations of the response, the decision was made to try estimating three models using the advanced modeling framework; one model for each time of day at the partial segment:

- morning model for 7:00-10:59;
- midday model for 11:00-15:00;
- afternoon model for 15:00-20:00.

In selecting the time period for each model, care has been taken not to make periods too long, but nevertheless long enough to capture both beginning and end of the rush hour if possible; or either of those for longer (i.e. afternoon) rush hour. Each of these models has a slightly different set of covariates, as covariates with large p-values have been rejected (one at a time) and model refitted with a reduced set when necessary. The detailed results are reviewed in the following section.

### **5.4.3 Model Diagnostics and Results**

From the previous section, three models have been estimated on the partial road segment, each for a different time of day. Because the number of observations for each model is very large,

Q-Q plots, studentized residual plots, Cook’s distance plots, or any other standard plotting diagnostics to test the presence of heteroskedasticity and outliers are infeasible. For this reason – and also for the fact that outliers are not quite possible since the source variable for the response is bounded and covariates are designed and controlled by us – we have decided to trust the theoretical foundations of linear correlation modeling laid out by Bartlett (1993) and tested as described in Asuero, Sayago and González (2006).

The three models presented in Table 4 below are using the advanced building concept, as it offered better fit than its simple counterpart at any stage of development described previously.

**Table 4 Summary of the Final Models Estimated for Partial Segment of I-10 E**

Variable	<u>MORNING</u>		<u>MIDDAY</u>		<u>AFTERNOON</u>	
	Estimate	t-statistic	Estimate	t-statistic	Estimate	t-statistic
intercept	1.146	893.62	1.201	961.52	1.023	1199.53
miles	-0.3859	-121.8	-0.4229	-137.09	-0.2075	-110.57
timediff	-0.006456	-70.5	-0.008089	-325.99	-0.0068845	-166.88
miles^2	0.007377	57.42	0.007032	56.4	0.008900	103.53
timediff^2	1.314E-05	130.16	1.449E-05	147.41	1.409E-05	326.85
miles*timediff	0.001616	40.03	0.002297	73.95	0.001147	59.32
miles*state	0.05589	59.43	0.02229	24.35	0.02027	30.33
timediff*state	-0.0002632	-9.49	-0.0003689	-15.1	-0.0004872	-34.01
miles*timediff*state	-0.0002045	-16.97	7.099E-05	6.24	-7.917E-05	-11.24
miles*lanes	0.03593	54.31	0.04913	76.29	0.009453	22.61
timediff*lanes	-0.0004985	-26.67	-----		0.0001714	18.7
miles*timediff*lanes	-9.987E-05	-11.67	-0.0002986	-45.52	-0.0001322	-30.23
miles*RN_dummy	0.01206	26.48	-0.01510	-34.06	-----	
timediff*RN_dummy	0.0001859	14.13	-0.0002644	-20.88	-0.0004015	-57.04
miles*timediff*RN_dummy	-8.995E-05	-14.99	0.0001600	27.57	0.0001397	54.41
miles*NN_dummy	0.02336	22.91	-0.03120	-31.42	0.006763	9.95
timediff*NN_dummy	0.0004351	16.01	-0.0002547	-9.83	-0.0005855	-39.93
miles*timediff*NN_dummy	-0.0002060	-16.5	0.0002331	19.38	0.0001711	24.68

Please note that each of the above models offers a comparable or better fit to the raw data than the preliminary model has to the averages. However, when comparing the estimates to those

from Table 1 (which summarizes regression for the preliminary study), it is important to remember that the models above work on Fisher Z transforms instead of correlation directly.

Some of the error in all models comes from the fact that, when both *miles* and *timediff* are equal to 0, intercepts do not transform back to the correct value of 1. This is because all self-correlations were previously removed from the data sets to avoid program errors during Fisher transformation (division by 0).

Although coefficients for *miles* and *timediff* are all negative, the absolute values for those associated with *miles* are much stronger. Please keep in mind, however, that *miles* only varies by several units, while *timediff* is given in minutes and its effect should not be underestimated (minutes can vary from 0 to 240 or 300, depending on the model). Nevertheless, the negative effect of spatial distance is much stronger than that of temporal distance. For example – to use most extreme cases – the effect of 1 mile in the midday model is equivalent to 51 minutes, but only 28 minutes in the afternoon model. These numbers do include the correction of the quadratic terms. Agreeing with the results of the preliminary regression, positive signs of time-space interaction terms suggest that the rate at which correlation drops decreases with the increase of spatial and temporal distance.

Positive parameter for the interactions of spatial distance with the level of congestion – *miles\*state* – suggests that increasing average congestion level between two locations reduces the effect of spatial dimension (reminder: state changes from 0 to 1, when 1 means both locations are always unstable). This is quite on the contrary to the temporal dimension. Sign for the interaction of *state* with the time-space interaction term varies from model to model. In the overall effect (simulated), higher congestion increases the predicted response for all models.



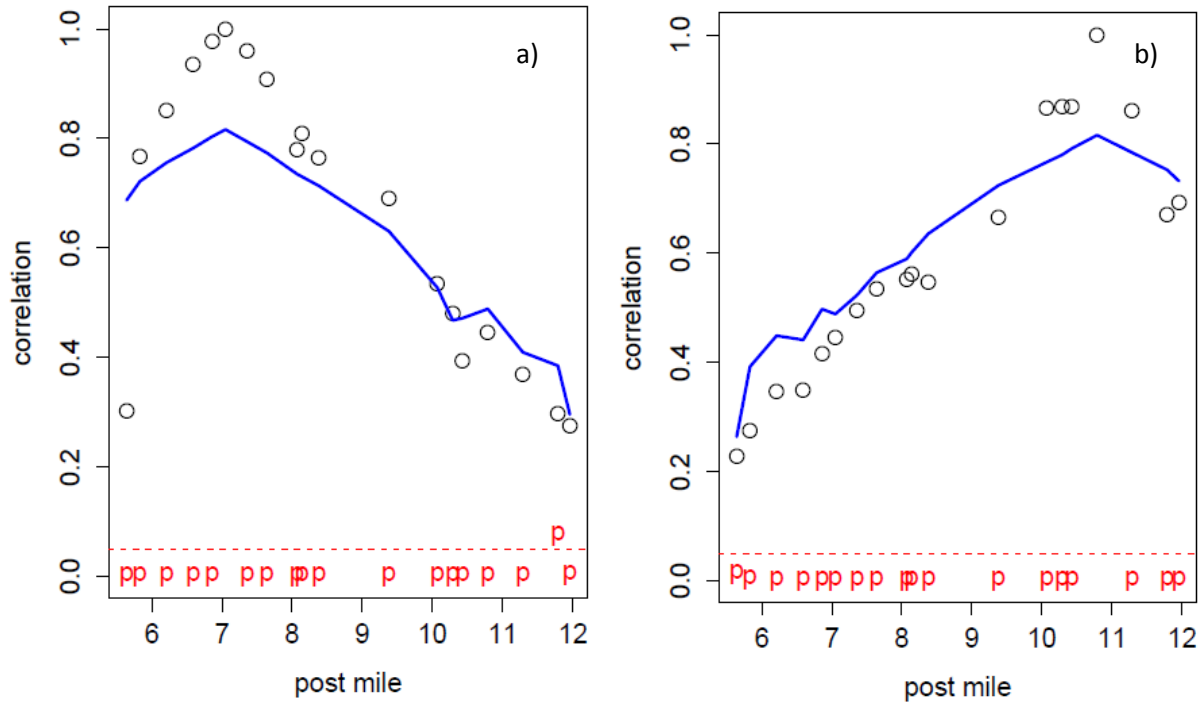
Average number of lanes – similarly to state – reduces the spatial dimension’s effect when there are more lanes available, but also brings more negative drop when combined with time-space interaction term. The effect of *lanes\*timediff* varies from model to model, and has been insignificant in the midday model. Simulation showed that in the overall effect, adding lanes increases the predicted response in both morning and afternoon models.

The overall effect of the ramp presence has been shown to slightly reduce or slightly increase response estimate in the morning and midday model, respectively, when compared to the situation when both locations have ramps (*NN\_dummy*=*RN\_dummy*=0). In both cases the effect was smaller when only one reference location had ramp, i.e. *NN\_dummy* induces higher overall change than *RN\_dummy*. The exception is the afternoon model, which seems to be unaffected by the ramp presence, at least in the simulated scenarios.

#### **5.4.4 Morning Model Evaluation**

Of the three models presented, one that is fitted to the morning data delivers the best fit. This is not surprising, as morning rush hour is much shorter than in the afternoon and trips made by morning commuters are typically more predictable and involve less variability than afternoon trip patterns. This ‘more-ordered’ morning trip pattern perhaps translates into more ordered and easier-to-predict correlation pattern. Morning model begins before and ends after the morning peak.

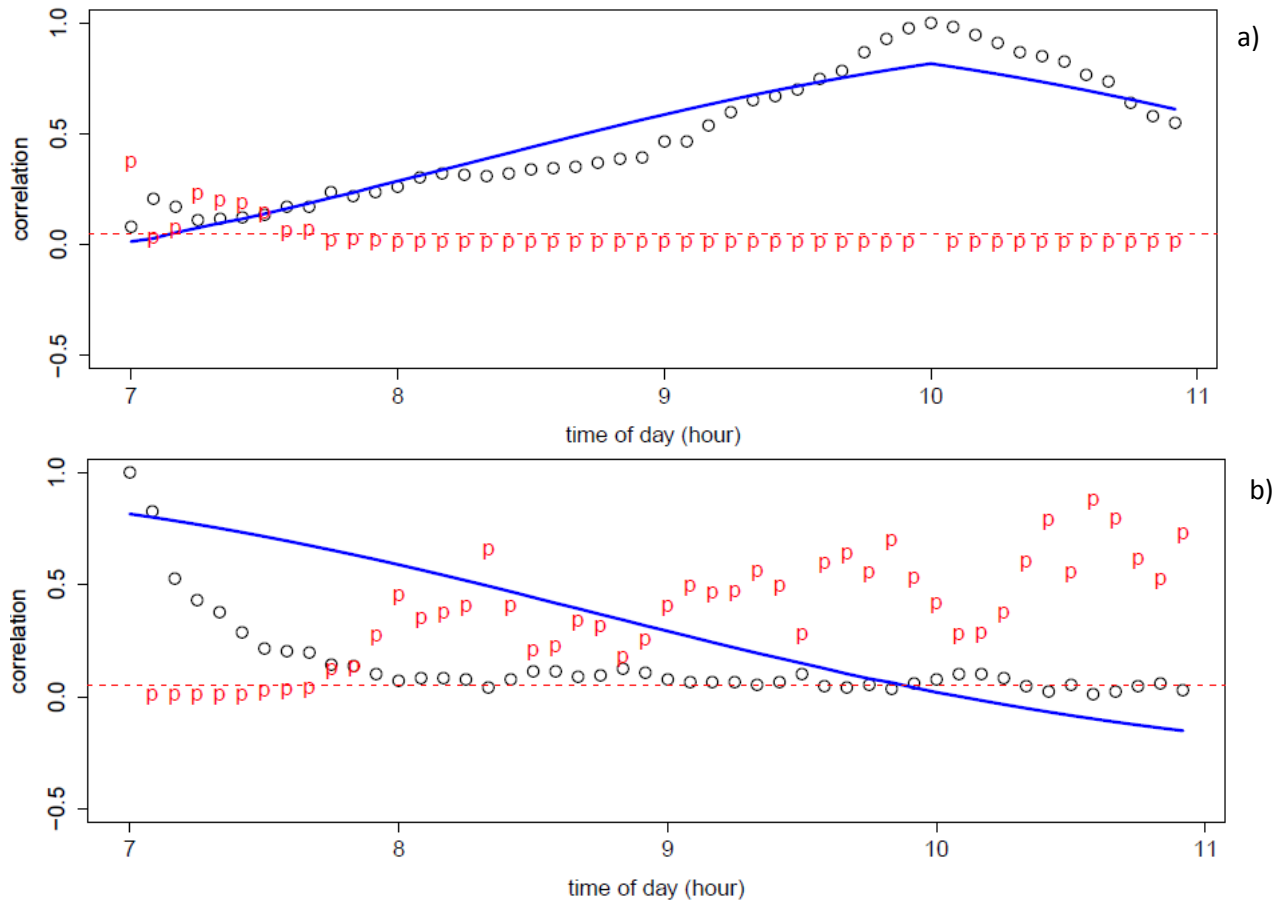
Figure 8 presents the observed spatial correlation patterns with indicated p-values and the plot of predicted correlation values. Spatial correlation plots show that morning model performs well to delineate the correlation trends. Corrections to the function from geometry changes can sometimes correctly predict the location, but not necessarily the shape of disturbance in the general trend.



**Figure 8 Morning Model: Observed vs. Predicted Spatial Correlation.**

**a) PM=7.05 at 9:00, b) PM=10.79 at 9:00**

Figure 9 provides insight into morning models' performance in the temporal dimension. Temporal correlation plots suggest not as even fit of the predicted function to the observed values; Figure 9 a) shows similar regression performance to that of the spatial dimension, but Figure 9 b) indicates significant prediction error. Nevertheless, the latter one is quite interesting as it shows the very steep drop of correlation. The correlation reaches (and remains at) near-zero values only minutes after the beginning of the morning rush hour; it shows that morning rush effectively filters out any dependencies of early morning with the rest of the day. The smooth shape of the temporal predictions is due to the fact that all geometry corrections remain constant when spatial dimension is set to 0.



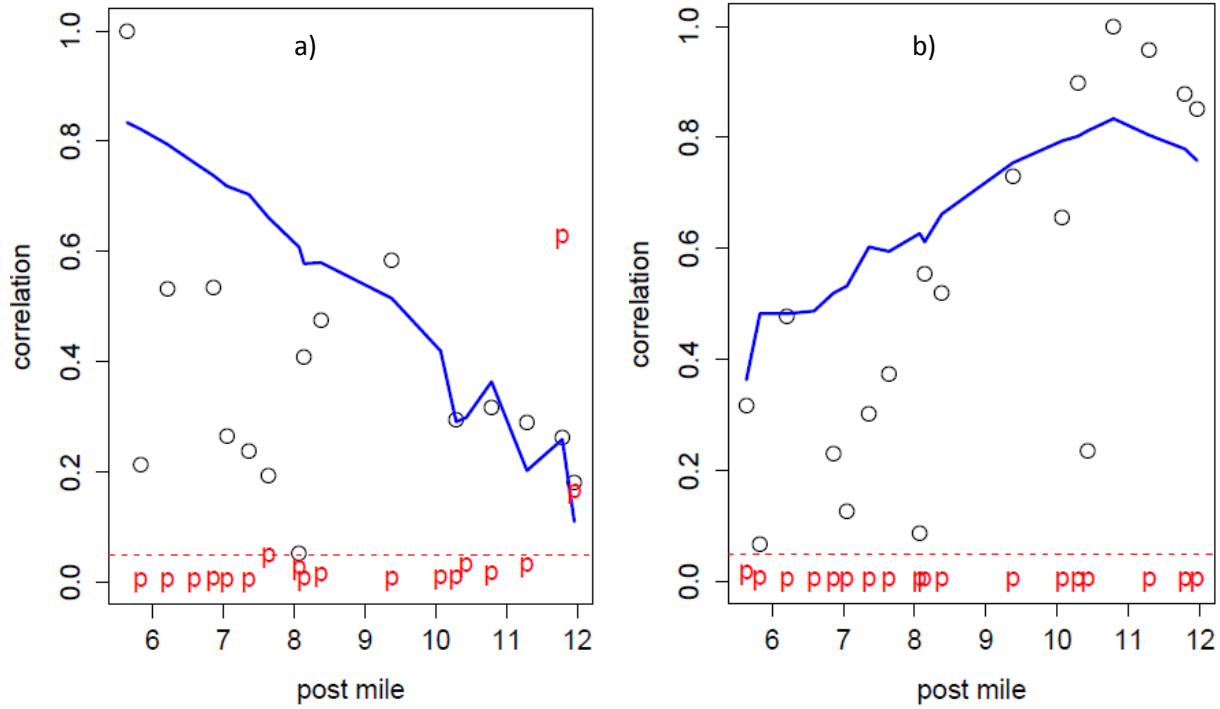
**Figure 9 Morning Model: Observed vs. Predicted Temporal Correlation.**

**a) PM=6.86 at 10:00, b) PM=8.07 at 7:00**

### 5.4.5 Midday Model Evaluation

Midday model's performance places it in the middle between morning and afternoon. It encompasses most of the period between the two rush hours and ends well after the beginning of the afternoon peak.

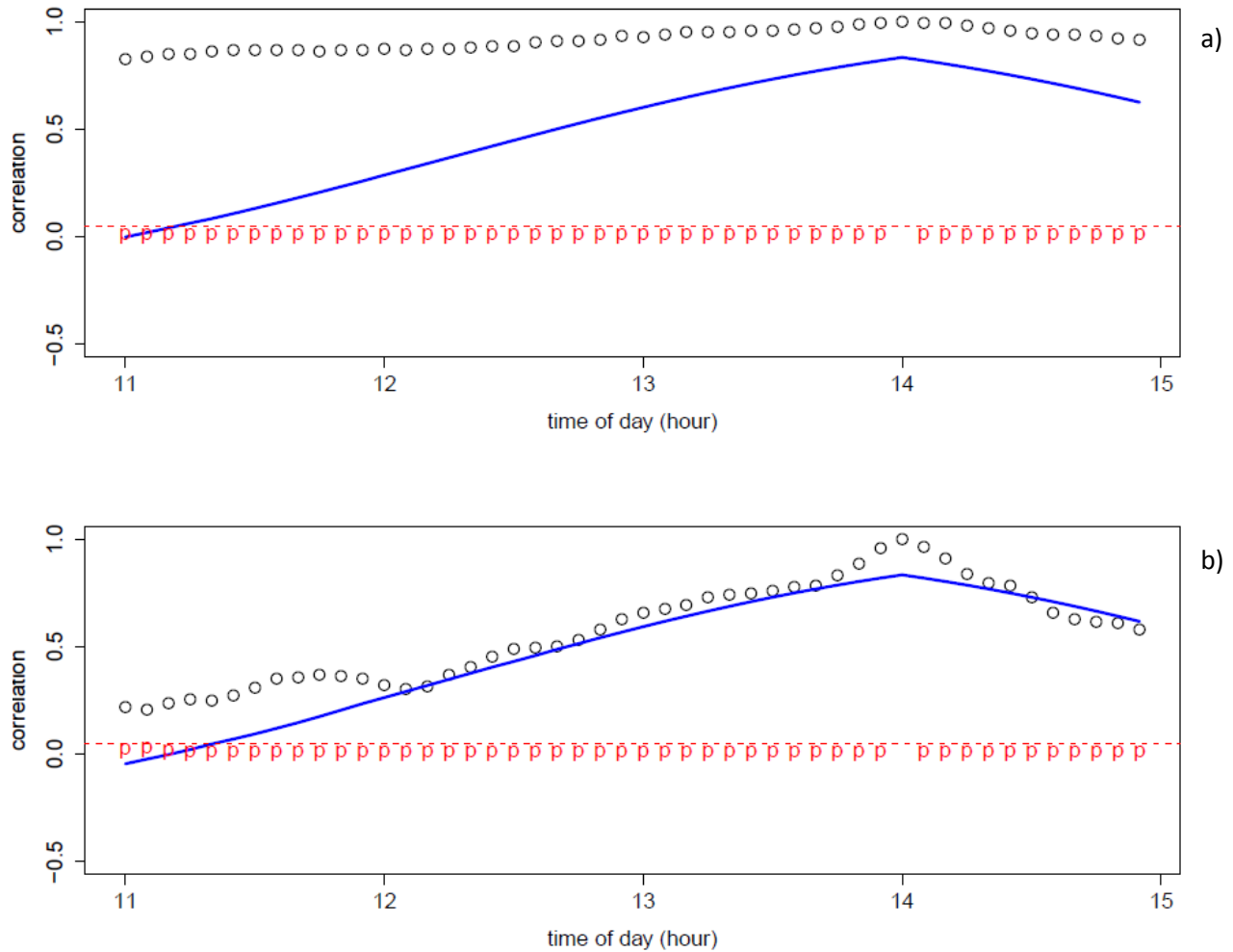
Figure 10 depicts observed vs. predicted spatial correlation patterns for midday period. Spatial correlation at off-peak hours (noon) seems to follow a less ordered, more scattered pattern than during the morning peak. Model does not capture that variation, yet it predicts the overall trend quite well.



**Figure 10 Midday Model: Observed vs. Predicted Spatial Correlation.**

**a) PM=5.64 at 12:00, b) PM=10.79 at 12:00**

Figure 11 shows those from the temporal perspective. Both temporal correlation plots are done for the same reference time, but the shape of the observed correlation and how well the predictions fit are very different. In Figure 11 a), the observed correlations are all very high and nearly constant. The regression line looks quite irrelevant in this case; however, during the same reference time but at a different location (Figure 11 b)), it performs exceptionally well. The suspected explanation for constant correlation might be that this particular location experiences little variation in speeds no matter the time. Confirmation of this hypothesis needs further investigation in the underlying data. It needs to be mentioned, however, that observed temporal correlation at this particular location had similar shapes also during morning and afternoon hours.



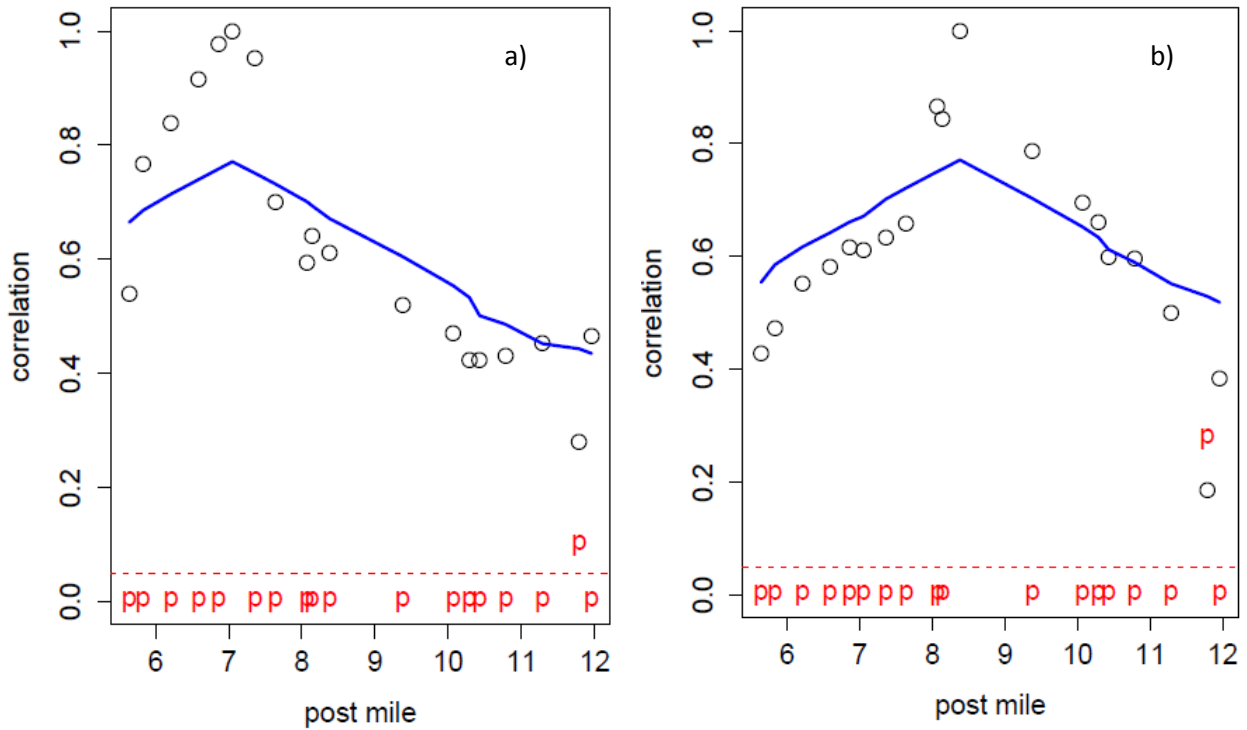
**Figure 11 Midday Model: Observed vs. Predicted Temporal Correlation.**

**a) PM=5.64 at 14:00, b) PM=8.07 at 14:00**

#### **5.4.6 Afternoon Model Evaluation**

Afternoon model delivers the worst fit of the three, but it is still roughly at the same level than the fit of the preliminary model. Using analogy to the morning model and typically more chaotic nature of the afternoon peak, it is believed that noise in the afternoon trip patterns is the underlying cause of the loss in this model's fit.

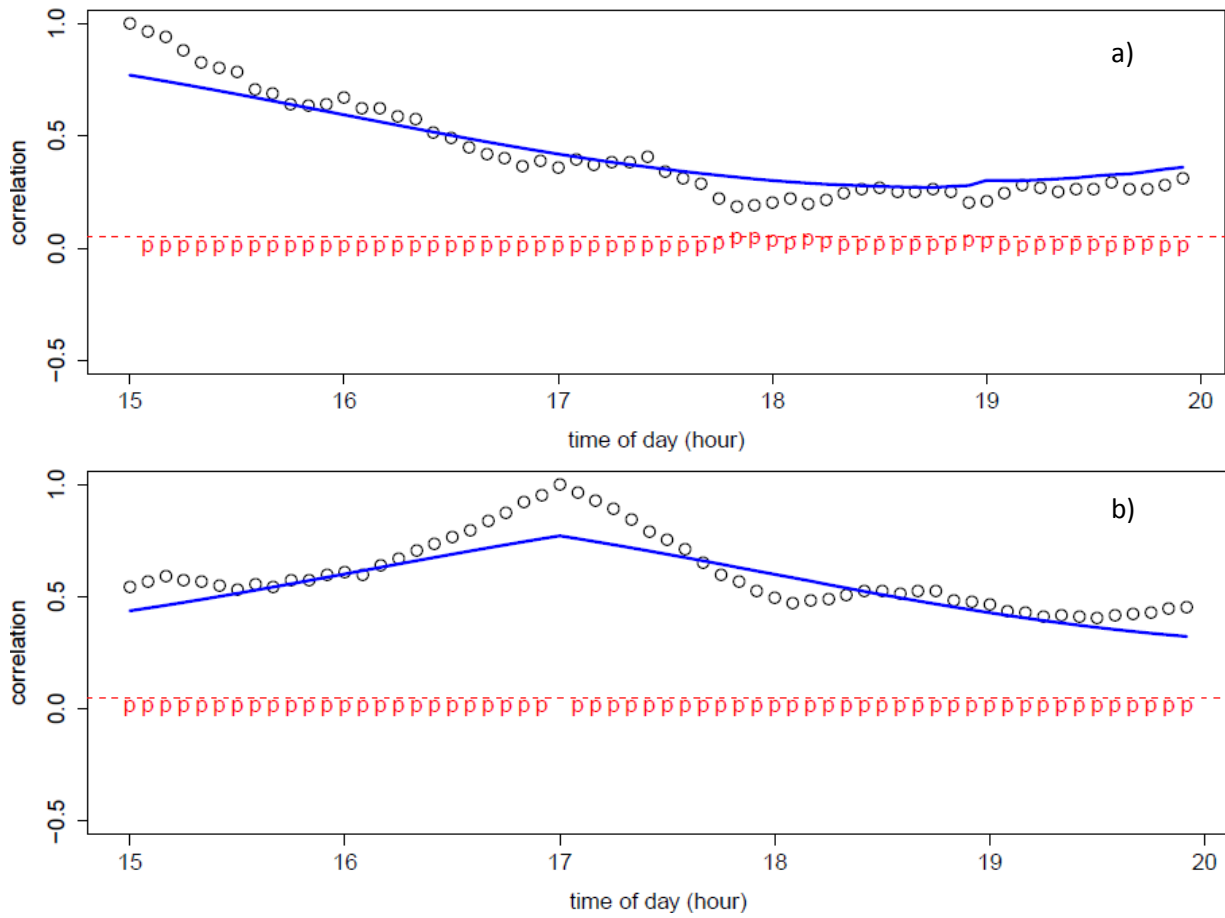
Figure 12 depicts observed spatial correlation vs. correlation predicted with the afternoon model. Figure 13 puts them into the temporal perspective.



**Figure 12 Afternoon Model: Observed vs. Predicted Spatial Correlation.**

**a) PM=7.05 at 17:00, b) PM=8.38 at 17:00**

The afternoon model delivers worst overall fit, but – as plots included in Figures 12 and 13 show – it can provide a reasonably accurate prediction of correlation in both spatial and temporal dimension. Virtually entire horizontal axis scale in Figure 13 is showing periods classified as afternoon rush hour – it starts well before 15:00 and ends just before 20:00. While the rate of decrease in Figure 13 a) is much more gradual than in Figure 9 b) – the periods in a peak nevertheless lose all the dependency not just to off-peak periods, but also to other periods within the very peak.



**Figure 13 Afternoon Model: Observed vs. Predicted Temporal Correlation.**

**a) PM=6.86 at 15:00, b) PM=10.07 at 17:00**

## CHAPTER 6

### CROSS-VALIDATION

#### 6.1 Background and Motivations

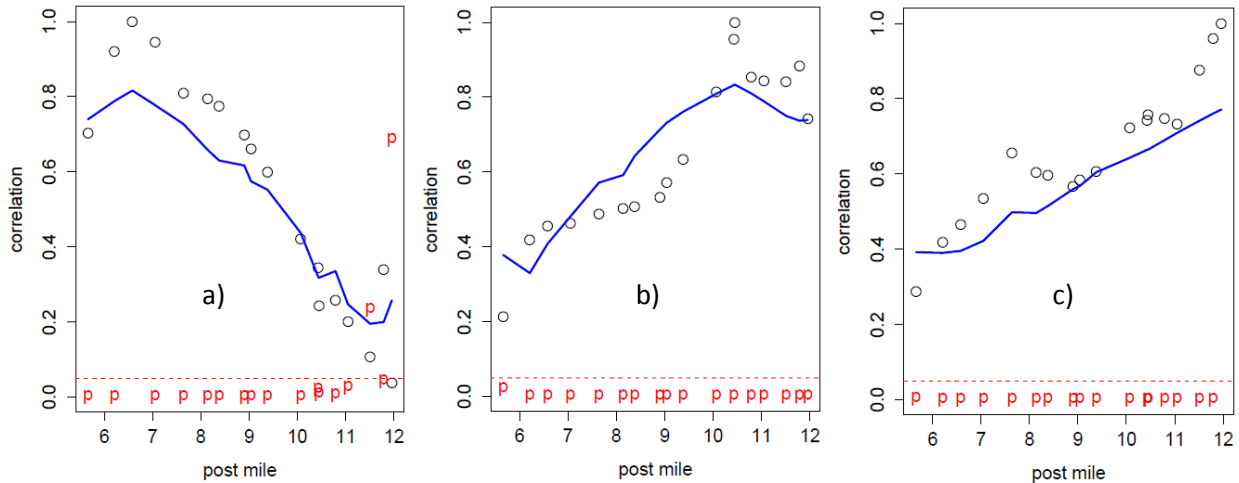
The regression models described in the previous chapter provide a satisfactory fit to the data taken from a single location. To answer the question whether correlation is location-specific – as suggested in the preliminary study – the models need to be transferred to other freeway settings with different traffic pattern and geometrical arrangement. Cross-validation then aims to test the robustness of our models but also to provide insight into how correlation patterns are dependent on location. Two distinctive new settings are discussed in the following sections.

#### 6.2 Location 1: I-10 Westbound

The first choice for location to cross-validate the models is to use the opposite direction of the freeway between the similar range of post miles. Thus, a segment of I-10 W between PM 11.96 and PM 5.66 is chosen. It is likely that traffic patterns are different from the corresponding stretch of I-10 E, since this is the outbound link from downtown Los Angeles. Per Appendix 1, road geometries between the two directions are comparable when one notices the frequent ramps, but by no means are the geometries mirror reflections of each other. Number of lanes varies differently on westbound and ramp arrangement is different.

Figure 14 shows observed vs. predicted correlation of all three models in the spatial dimension; Figure 15 shows observed vs. predicted temporal correlation. All predictions are done using the parameter estimated on I-10 E, applied to I-10 W covariates and transformed back to correlation domain.

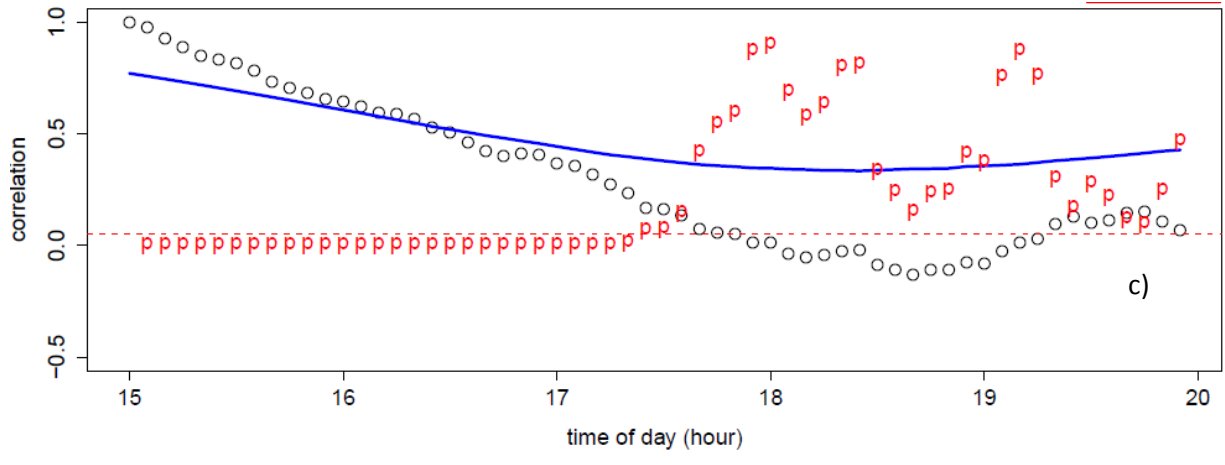
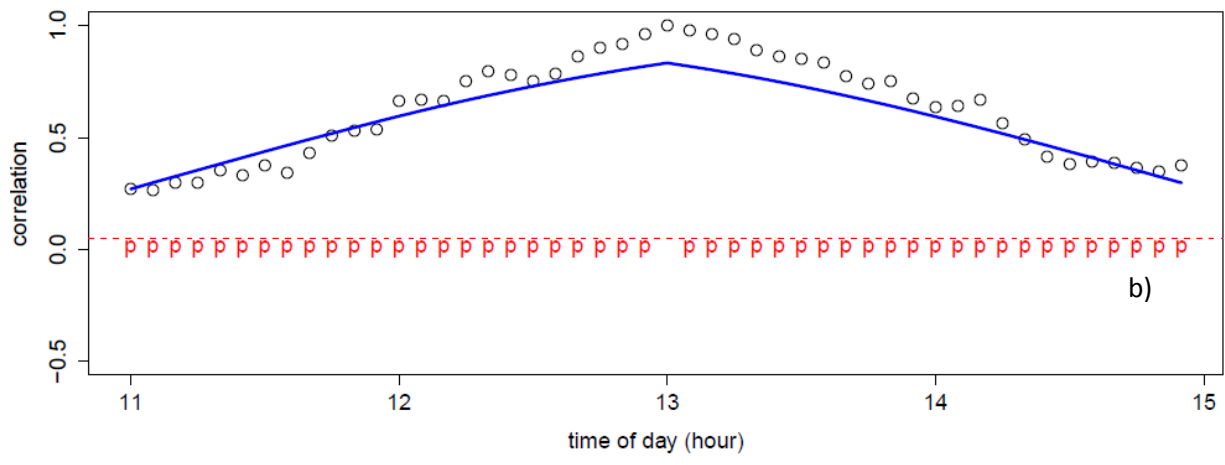
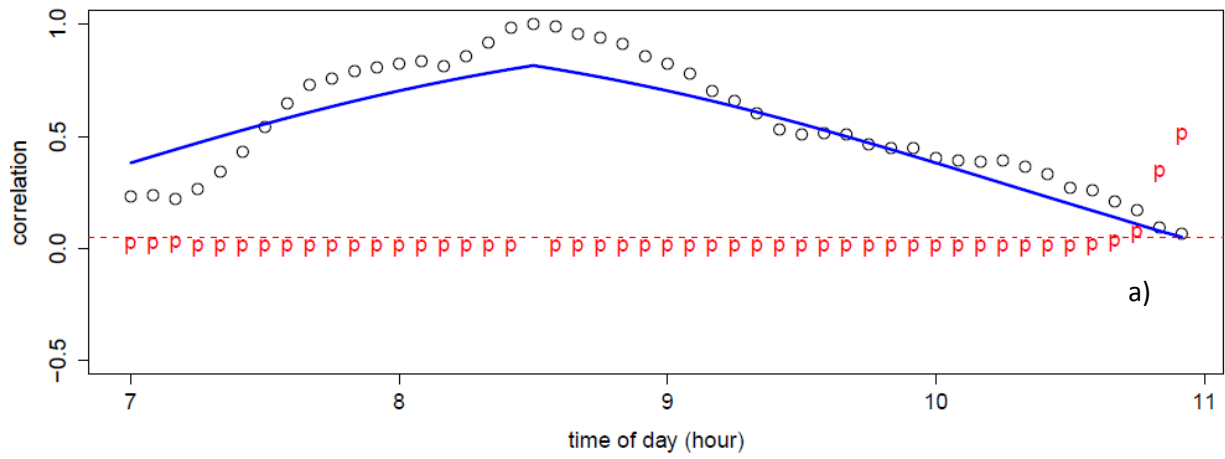




**Figure 14 Spatial Correlation Cross-Validation, Observed vs. Predicted (I-10 W).**

**a) Morning: PM=6.58, 9:00, b) Midday: PM=10.45, 12:00, c) Afternoon: PM=11.96, 17:00**

The spatial correlation plots indicate that both morning and midday models performed exceptionally well in fitting to the observed values. Afternoon model suggests slightly more prediction error and perhaps underestimates most observations, but it still provides a good approximation. The comments for Figure 14 can as well be applied to Figure 15, as again morning and midday models deliver good prediction accuracy in contrast with the afternoon, which first underestimates, then overestimates correlation.



**Figure 15 Temporal Correlation Cross-Validation, Observed vs. Predicted (I-10 W)**  
**a) Morning: PM=11.96, 8:30, b) Midday: PM=10.79, 13:00, c) Afternoon:**  
**PM=9.38, 15:00**

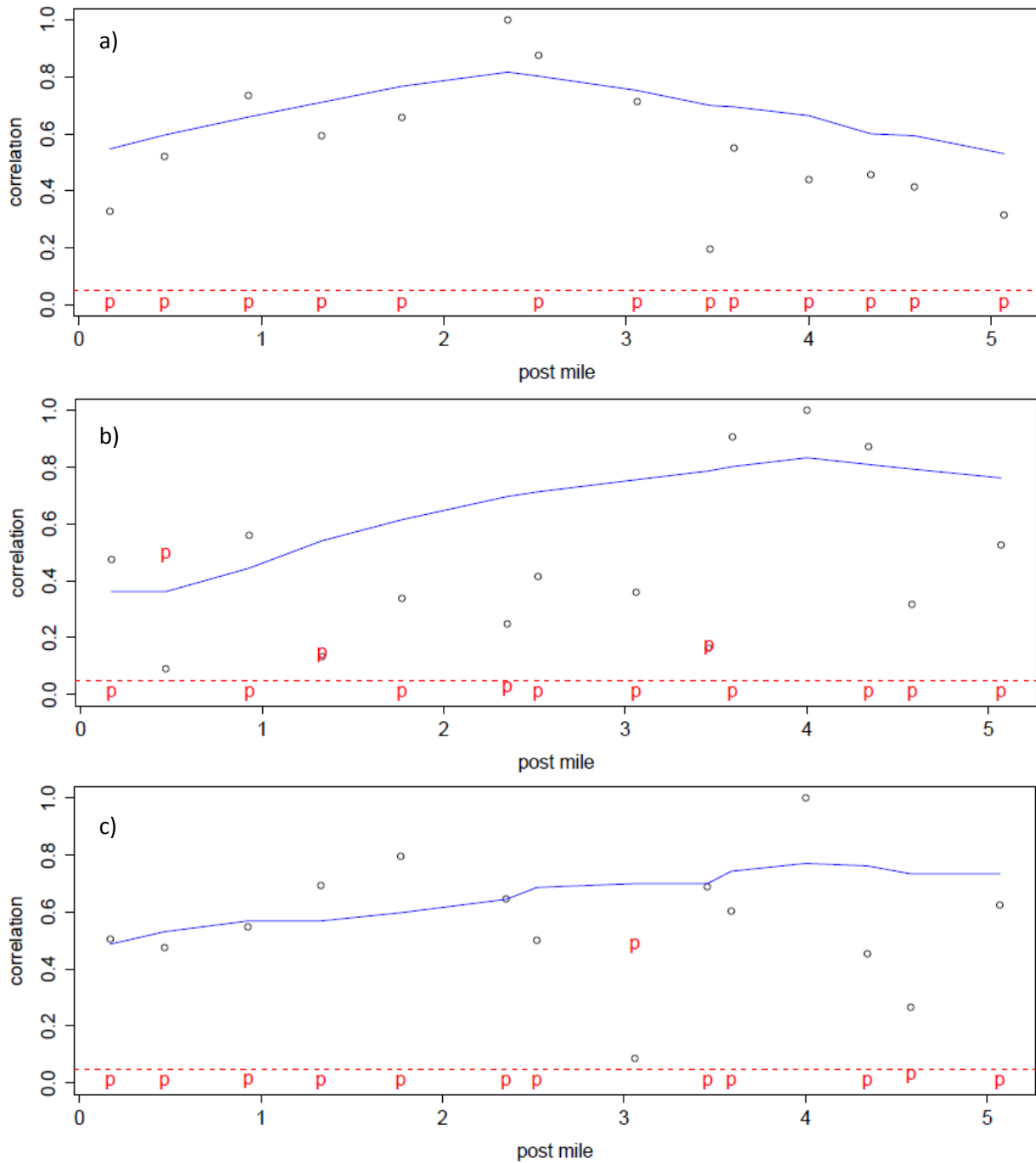
### 6.3 Location 2: I-10 Eastbound (Beginning)

The second choice for cross-validation with the data at hand is to test the models on the eastbound stretch that has not been used for estimation. The overall segment is divided to provide a split into two distinctive links with different geometrical characteristics. Therefore, this beginning stretch spans over 14 detector stations and ends before a major interchange with Venice Boulevard.

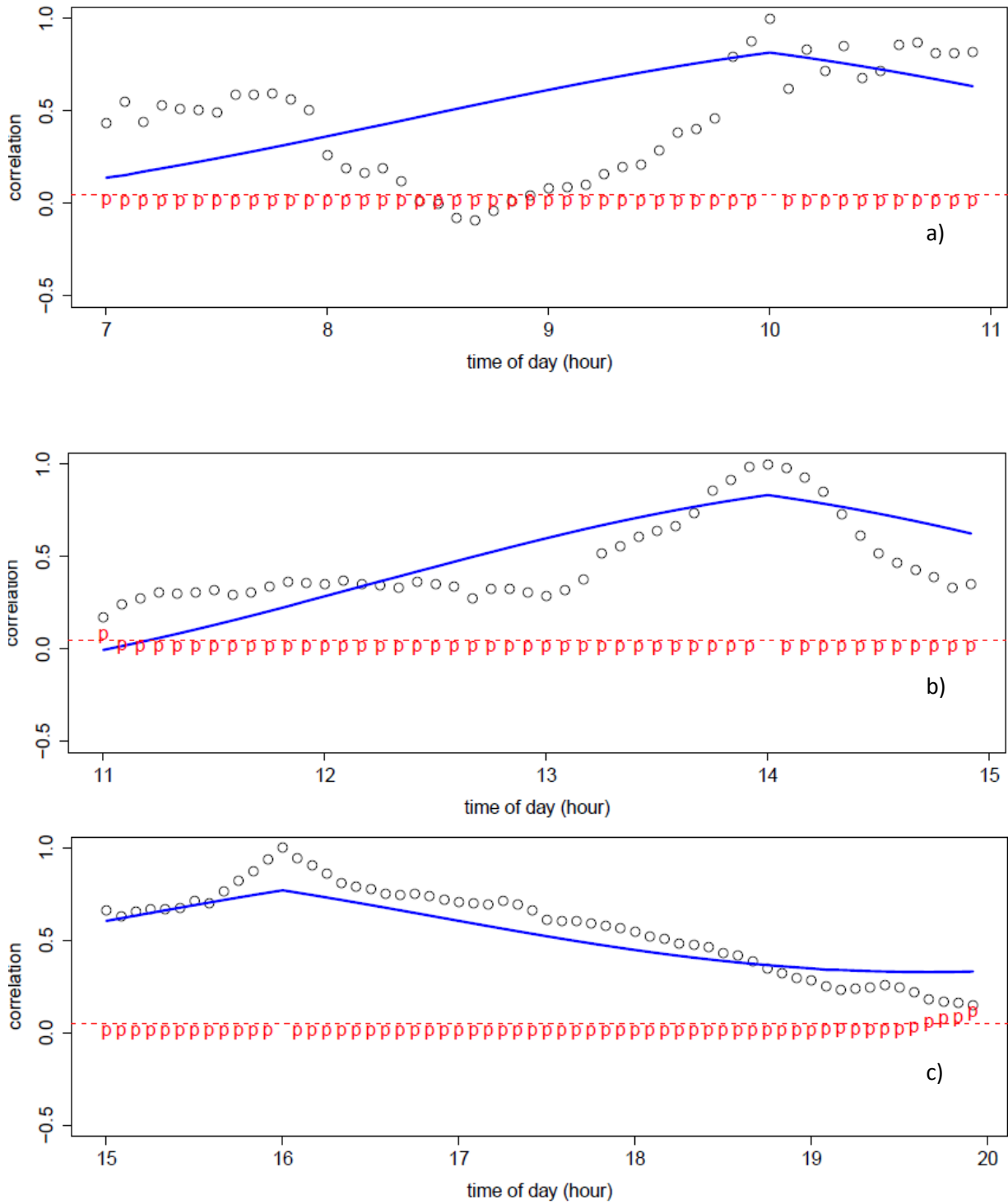
Figure 16 shows observed vs. predicted spatial correlation plots for all three models. In this new setting, all three models suggest a reduced performance. Still, each model retains the ability to predict the overall trend of correlation. Observed correlation coefficients tend to be more scattered compared to the other two investigated locations, and this increased variability can be noticed during each time of day.

Figure 17 shows the performance of models' predictive accuracy in the temporal dimension. Morning and midday plots agree with the comment regarding the spatial plots in that the correlation variability increased and that models are incapable of capturing all of the variation. Morning plot is additionally interesting because it may be seen as a complement to the morning plot shown in Figure 9 b): now, the reference period is after the morning rush hour, but still the sudden drop to zero is observed. This time, however, the correlations 'recover' and periods roughly before 8:00 retained some correlation to the reference. The afternoon temporal correlation plot is different from morning and midday, as its variability is significantly smaller, which works with benefit to the afternoon model's fit.

All of the comments to the figures regarding model fits pertain only to the situations on the presented plots and should not be generalized. Overall performance of each model at each location is discussed in detail in the following section.



**Figure 16 Spatial Correlation Cross-Validation (I-10 E, beginning). Observed vs. Predicted. a) Morning: PM=2.35, 9:00, b) Midday: PM=4.00, 12:00, c) Afternoon: PM=4.00, 17:00**



**Figure 17** Temporal Correlation Cross-Validation (I-10 E, beginning). Observed vs. Predicted. a) Morning: PM=0.17, 10:00, b) Midday: PM=1.77, 14:00 , c) Afternoon: PM=1.77, 16:00

## 6.4 Cross-Validation Summary

For proper assessment of the general predictive performance of each model, common and objective measures are needed. Please note that  $R^2$  can be interpreted as adjusted  $R^2$  when the sample size used for model estimation is very large. (Models presented in Section 5.4.3 are based on sample sizes at least  $4.2 \cdot 10^5$ , exact values depending on the model.)

Table 5 lists statistics associated with each model at each location.

**Table 5 Predictive Ability of the Regression Models on Three Independent Locations**

MODEL	Statistic	Est. Sample	Cross-validation at:	
		I-10 E (PM 5.64 - 11.96)	I-10 W (PM 11.96-5.66)	I-10 E (PM 0.17-5.04)
MORNING	bias	-0.01466	0.07839	-0.1417
	standard error	0.1600	0.1744	0.2587
	R square	0.6501	0.6198	0.4956
MIDDAY	bias	-0.01408	0.0186	-0.1623
	standard error	0.1437	0.1418	0.2749
	R square	0.6189	0.6345	0.4592
AFTERNOON	bias	-0.01200	-0.04899	0.05352
	standard error	0.1323	0.1720	0.1501
	R square	0.5405	0.4192	0.4980

Per Table 5, the capability of the morning model to correctly predict correlation on the westbound direction is about the same level as it achieved on its own sample, although it is perhaps slightly more biased; standard error remains roughly unchanged. When the same model is applied to the beginning stretch of the eastbound direction, all performance measures indicate a reduction of performance. Nevertheless, the model is still able to explain approximately half of the variance in the observed correlation.

Cross-validation of the midday model at the westbound sample indicates the same level of very low bias and standard error as at its estimation sample, with a slight increase in  $R^2$ . While delivering an impressive result at the opposite direction, the midday regression experiences the largest loss of predictive performance at the second cross-validation location.

The afternoon model achieves the smallest  $R^2$  when fit to the observed data in its estimation sample is assessed, but its bias and standard error are at the same time the smallest of all models. Although cross-validation to the westbound direction significantly reduced the fit, the afternoon model experiences the least performance loss when applied to the beginning stretch of the eastbound direction. At both cross-validation locations, it retains the very low level of bias and standard error.

Each model has a slightly different behavior when applied to independent cross-validation samples; however, all retain 75% or more of their original predictive ability. The developed regression models are capable of explaining significant percentage of variance in the observed correlation and the results show that these models can be transferred to new locations with some success.

## CHAPTER 7

### OTHER MODELING STRATEGIES

Models developed and validated in previous chapters offer the ability for correlation prediction; however, their predictions carry some error. The amount of the error varies and depends on both a model and its application. As mentioned on many occasions in sections pertaining to model building procedures, we have discovered a tradeoff between the number of situations to which a model can be applied and model's predictive performance. In other words, the more general the model's designed application, the worse fit it can deliver.

Three time-of-day specific models that we developed are based on data which were scaled down from the original dataset. Further split of the day or the segment into shorter lengths – though it might improve the fit – would generate even larger family of models and would probably make the application process impractical. Thus, we decided to look for a more careful choice of readings from the existing datasets with an objective to improve the fit of our models while retaining the general range of their applications.

Consider a traveler on freeway using the en-route traveler information system to advise his or her route choice decisions. If the information for traveler is updated with each ramp he or she passes, the driver can use such information to decide whether to continue travelling on freeway or exit at the next ramp. Providing information that is too far in advance may not be useful to some travelers; too much information at a time may also be too distracting and difficult to process while driving. Therefore it is possible that, in models we have discussed before, the ability to capture sudden drops or rises in correlation may be somewhat diluted by the models' attempts to provide information that is difficult to actually process. We decided to explore



building models that use the pool of predictor variables discussed in Section 5.4.1 and that follow the general model architecture outlined in Section 5.4.2.4; at the same time, datasets are heavily edited.

We propose to model correlation using only observations between adjacent stations. With a reference in station  $i$ , correlation can only be predicted at stations  $i-1$  and  $i+1$ . Under this strategy, temporal dimension is not affected and models are built again for morning, midday and afternoon. Each time of day is modeled using the full ‘advanced’ set of variables; predictors that generate largest p-values are iteratively rejected. Ultimately, a model for each time of day has been fitted to the station-to-station data. Table 6 summarizes fit statistics of the three models using this alternative approach.

**Table 6 Performance Indicators of Station-to-Station Models**

MODEL	Statistic	Est. Sample
		I-10 E (PM 5.64 - 11.96) STATION-STATION ONLY
MORNING*	bias	-0.004063
	standard error	0.06902
	R square	0.9774
MIDDAY*	bias	-0.003147
	standard error	0.05453
	R square	0.9843
AFTERNOON*	bias	-0.002852
	standard error	0.05344
	R square	0.9807

\* despite the same nomenclature, here presented models are not the same as previously described Morning, Midday and Afternoon models

All three station-to-station models are able to provide a near-perfect fit to the observed correlation. Bias is approximately an order of magnitude smaller than the smallest achieved using full sample; standard error is also 30-50% of the original best value.

Ability to achieve this level of fit consistently by all three time-of-day models undermines the idea to divide the day into different periods. It is possible that a single model can be estimated without scaling down the sample, yet delivering the fit to the data higher than multi-station models. It is unsure, however, how well station-to-station models would pass the cross-validation. Further research is needed to address some of these questions.

## CHAPTER 8

### CONCLUSIONS AND RECOMMENDATIONS

#### 8.1 Summary

A more thorough analysis that followed the initial simplified methodology resulted in the development of three regression models which model raw 5-minute correlation at the resolution of a station and take road geometry and variability of the state of traffic into account. Each model passes cross-validation onto two other settings and offers explanation of a significant portion of variance in correlation at multiple locations; however, as in Samaranayake, Blandin and Bayen (2011), the predicted values tend to have lower variability than observed.

All in all, the findings of the initial stage of the analysis are confirmed later in the paper – the primary factor for correlation is the spatial distance. In temporal domain, the slope of correlation trend tends to be more affected by the traffic state. Models agree that increasing congestion level creates a ‘correlation retaining’ effect – the rate of decrease with distance tends to be lower. Relatively small corrections to time-space effects are added when road geometry changes occur; however, it is believed that these parameters significantly improve the overall fit and play an important role – together with traffic state – when a model is transferred to a new physical setting. Cross-validation results indicate that at least 75% of each model’s predictive capability is retained when the models are applied to new physical locations. Thus, we can conclude with confidence that the models are transferrable.

## 8.2 Recommendations for Future Research

While assumption may be true that I-10 westbound direction, or that beginning segment of I-10 eastbound are independent from the latter stretch of I-10 E – it is also possible that it is not quite so. One way to check that would be to do a cross-validation onto a very different freeway, say I-880 in San Francisco Bay. Once successful, transferability of findings would be confirmed on even stronger grounds.

Although an improvement was anticipated, station-to-station sampling approach brought a fit to the observed correlation at an unprecedented level. Nevertheless, this thesis only scratches the surface of this approach, so this venue should be further investigated. Given the near-perfect predictions of these three models, future research should consider going back to whole-day data, or even 24-hour data. Possibly, much of the original fit could be retained while achieving simplicity of a single universal model. Such model should also be validated using an independent sample.

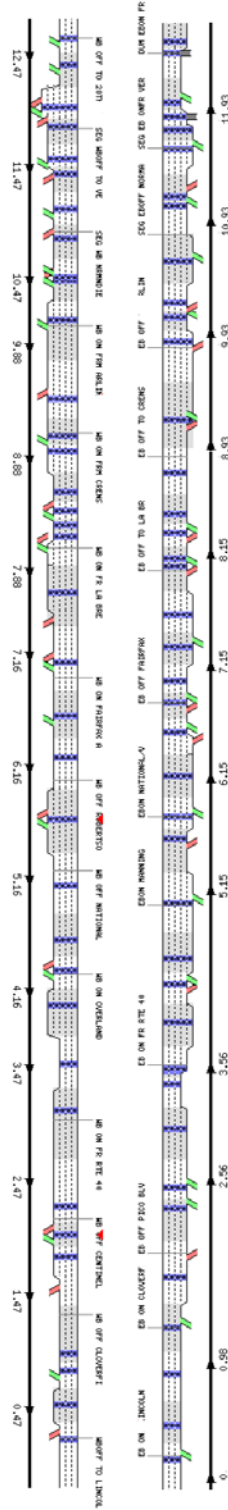
Lastly, this research only attempts to model correlation in a freeway setting, where traffic stream between consecutive exits can reasonably be assumed to be a continuous flow. Extending the methodology onto more chaotic arterials where there is no control for accessing or exiting the stream is expected to be a very challenging task in itself. Hopefully, this work will provide some useful foundation for extensions onto arterial roadways, and then onto entire transportation networks.

## APPENDICES

# APPENDIX 1

## LAYOUT SCHEMATICS OF INTERSTATE 10 BETWEEN PM 0 AND PM 12.50, BOTH DIRECTIONS

WESTBOUND – towards Santa Monica



EASTBOUND – towards Los Angeles

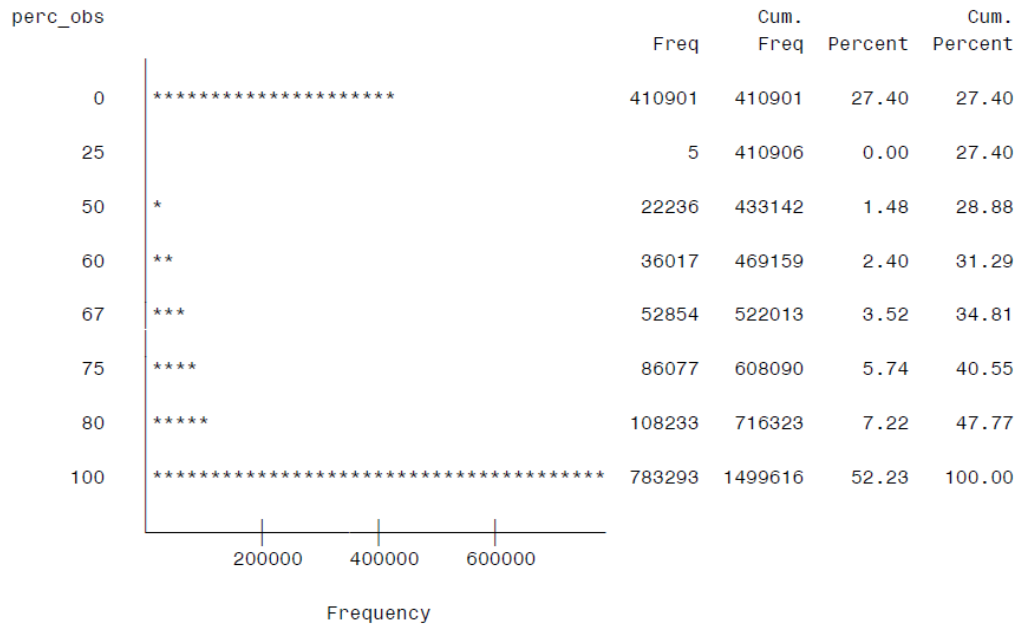
## APPENDIX 2

### DATA FILTERING PROCESS – FREQUENCY OF % OBSERVED DATA

#### Data Quality frequencies for I-10 E

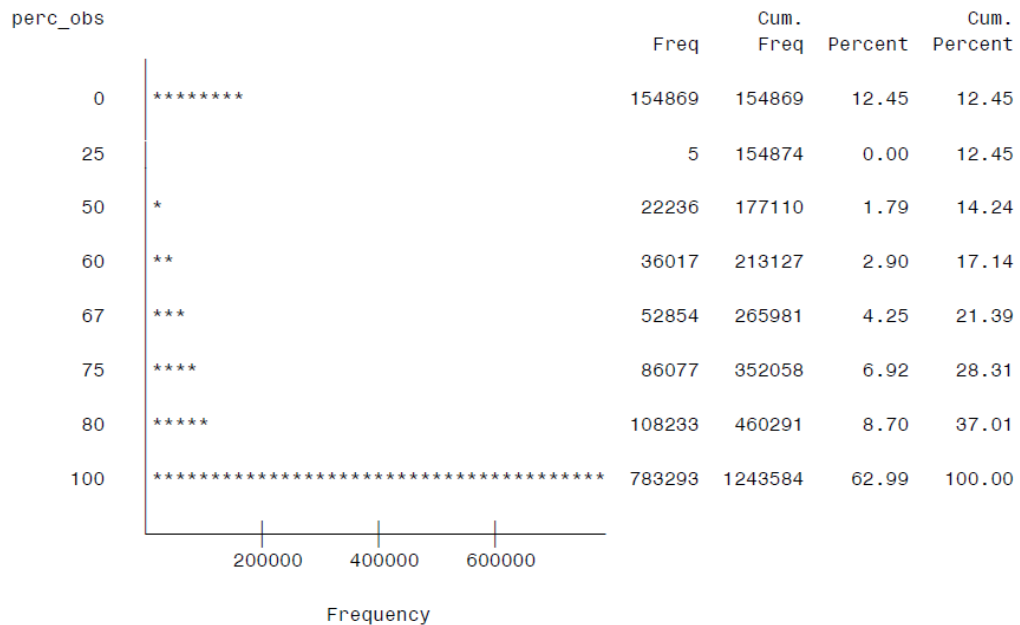
\*\*\*BEFORE REMOVING 0% STATIONS\*\*\*\*

The SAS System      08:09 Thursday, December 15, 2011 104



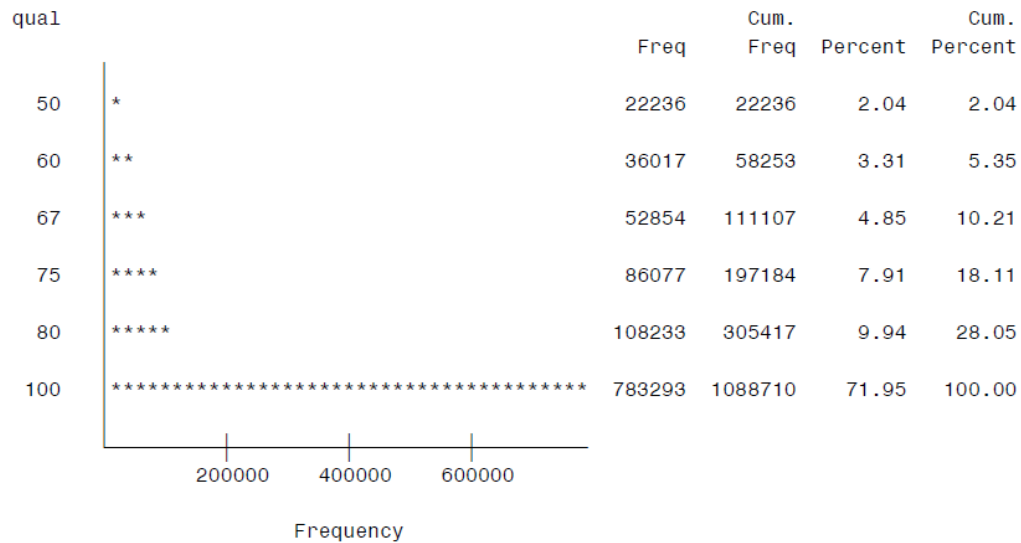
\*\*\*\*AFTER REMOVING 0% STATIONS\*\*\*\* (filtering level 1)

The SAS System      08:09 Thursday, December 15, 2011 110

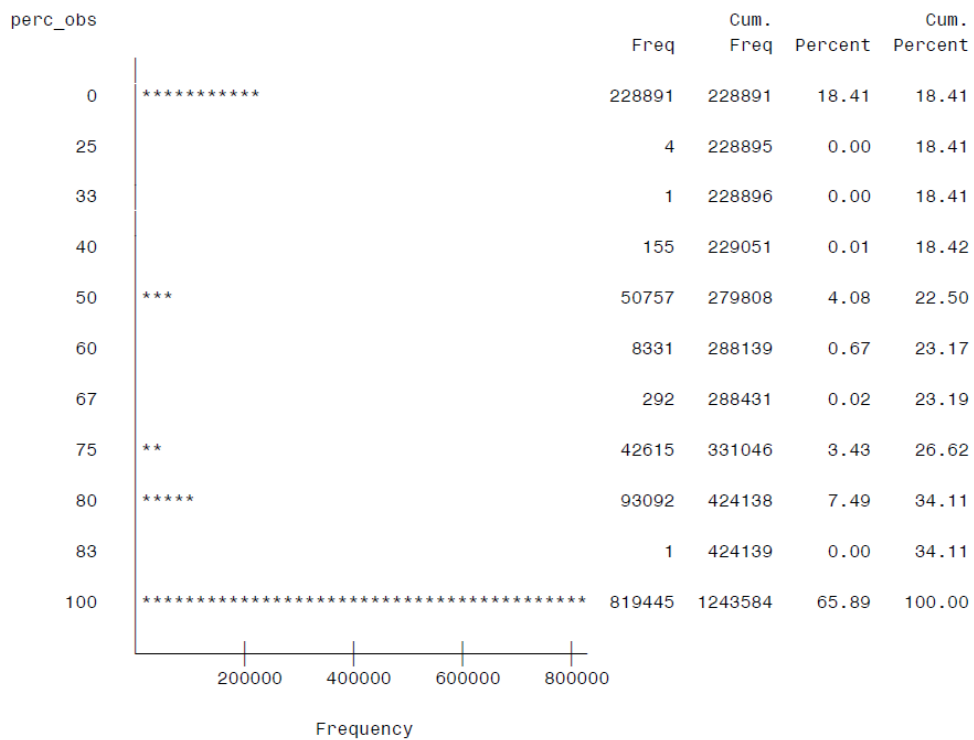


\*\*\*\*\*MISSING VALUES WHEN perc\_obs <= 50\*\*\*\*\* (filtering level 2)

The SAS System 08:09 Thursday, December 15, 2011 111

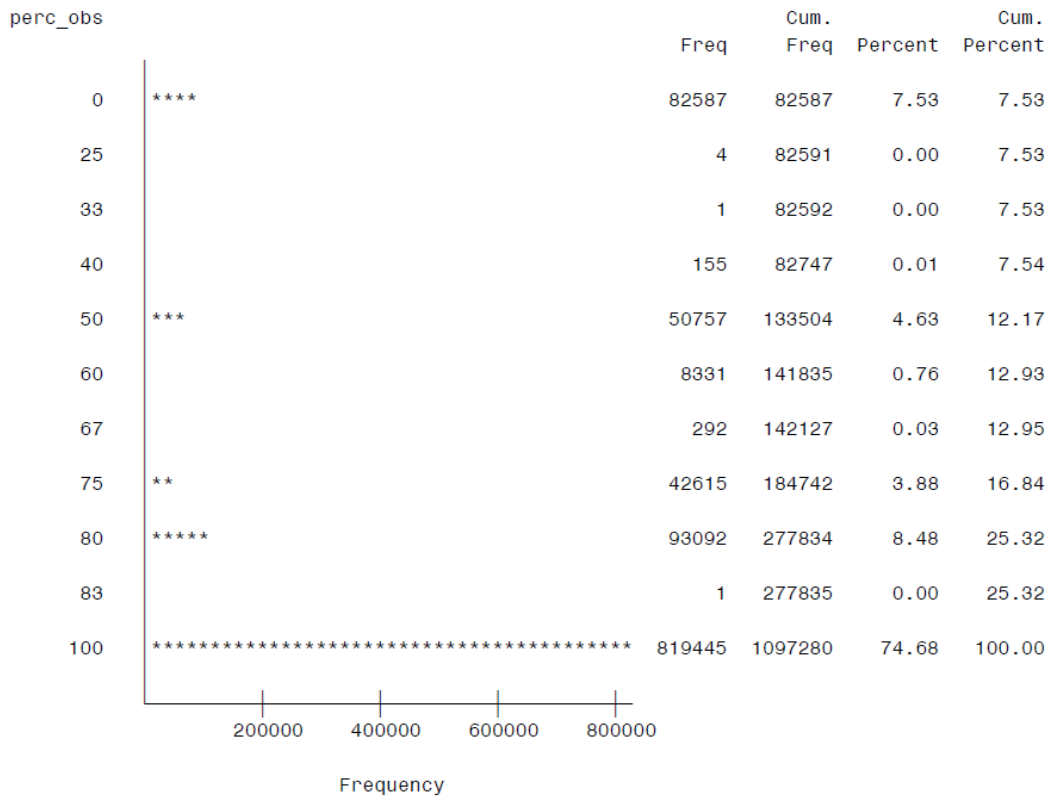


Data Quality frequencies for I-10 W \*\*\*\*\*BEFORE REMOVING 0% OBS STATIONS\*\*\*\*\*

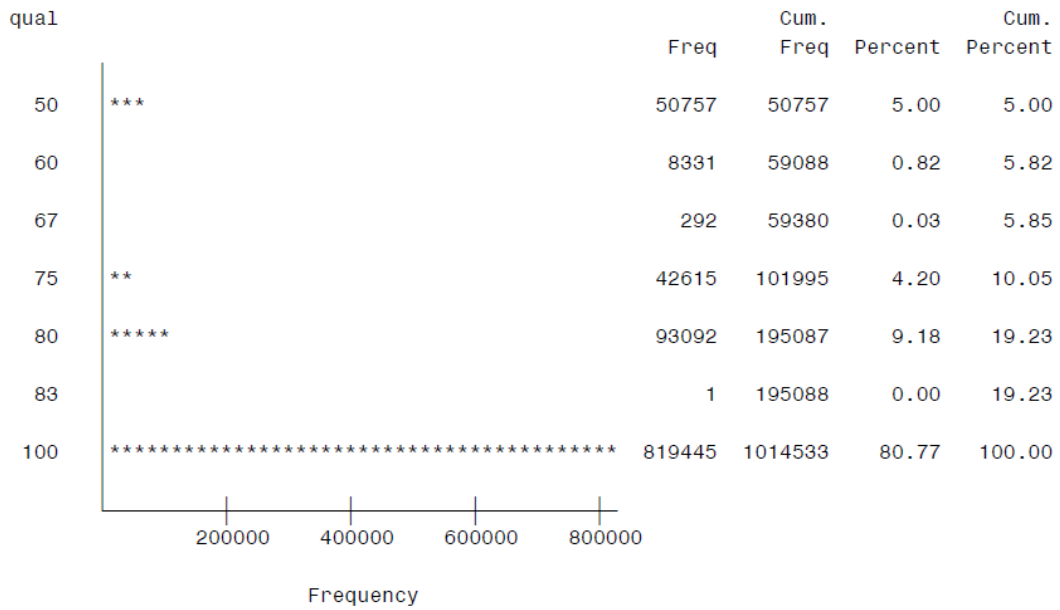




\*\*\*\*\*AFTER REMOVING 0% OBS STATIONS\*\*\*\*\*



\*\*\*\*\*FILTERING LEVEL 2 QUAL=% OBS IF % OBS>=50\*\*\*\*\*

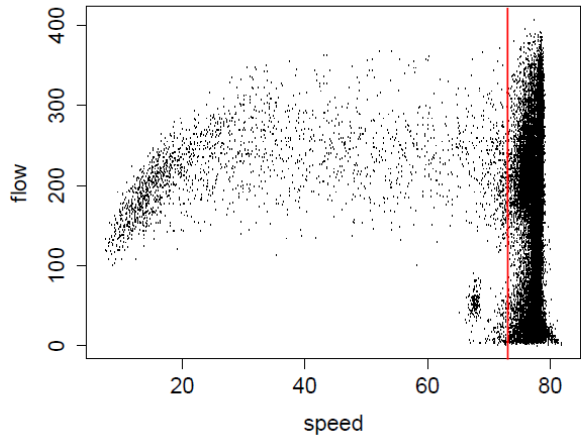


**APPENDIX 3**

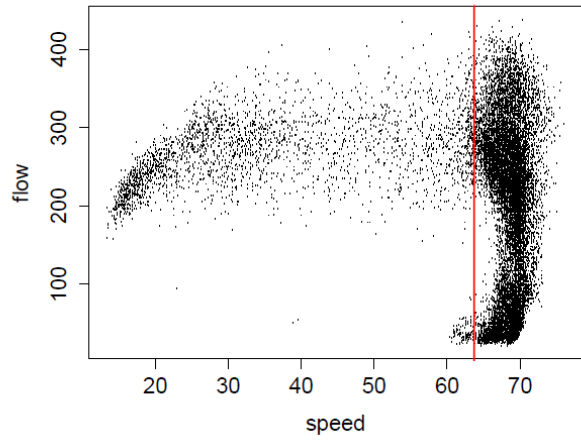
**SPEED-FLOW DIAGRAMS FOR I-10 E-W WITH CONGESTION THRESHOLDS  
INDICATED**

**A. EASTBOUND DIRECTION**

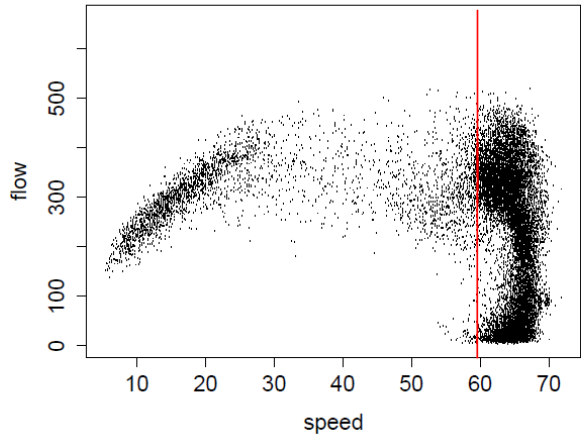
I-10 E, mile = 0.17



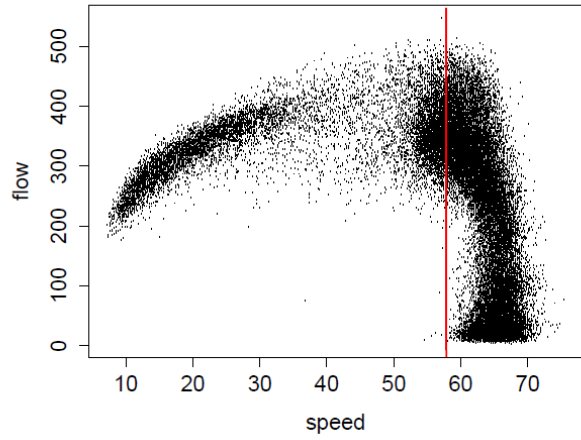
I-10 E, mile = 0.47



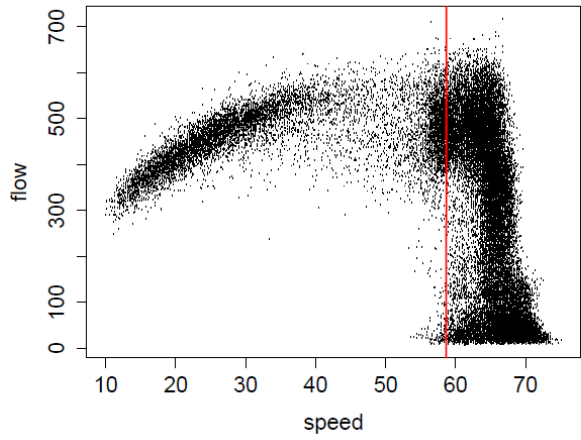
I-10 E, mile = 0.93



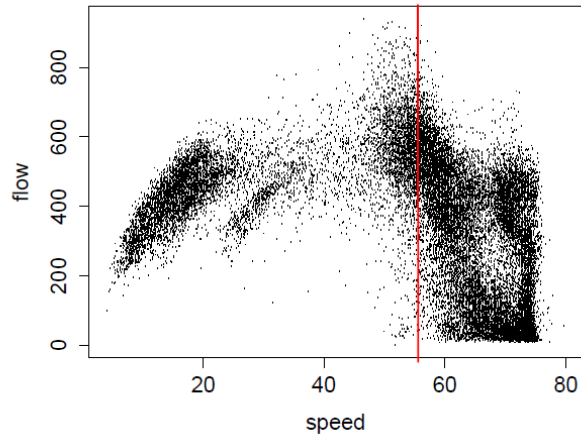
I-10 E, mile = 1.33



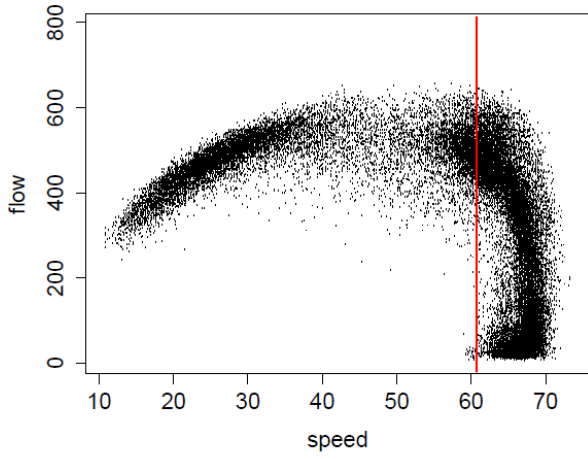
I-10 E, mile = 1.77



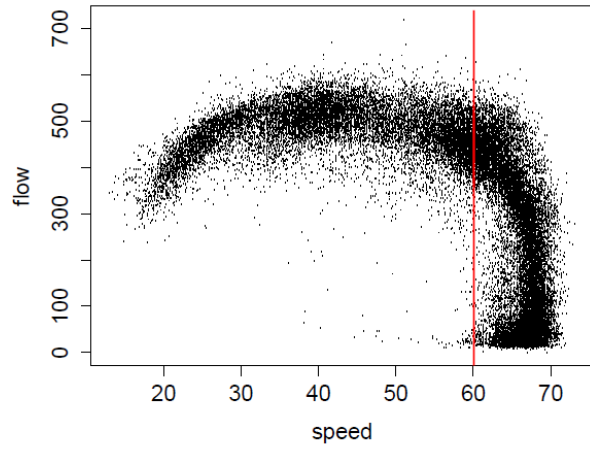
I-10 E, mile = 2.35



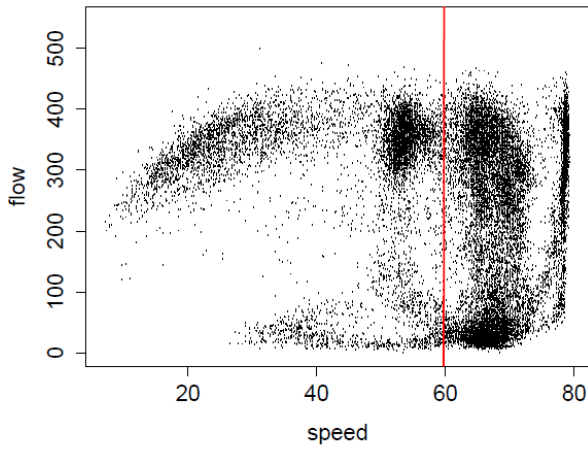
**I-10 E,mile = 2.52**



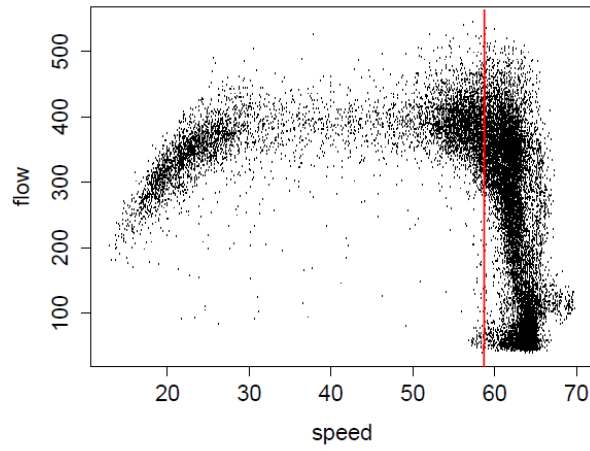
**I-10 E,mile = 3.06**



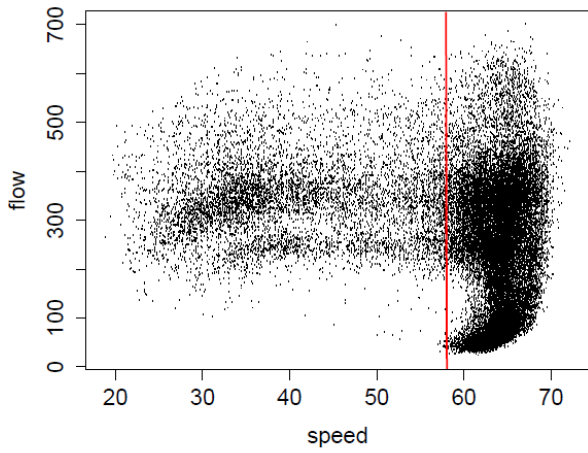
**I-10 E,mile = 3.46**



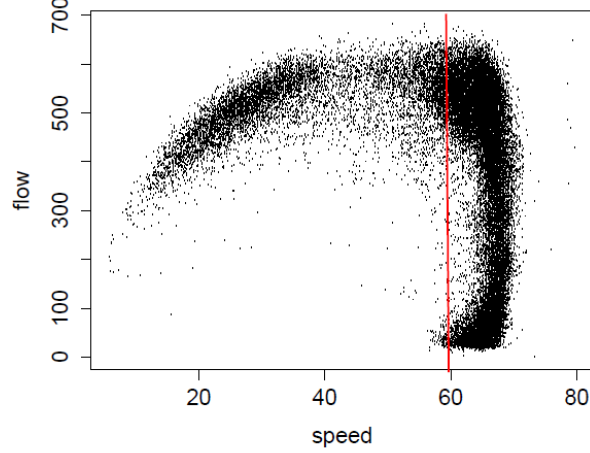
**I-10 E,mile = 3.59**



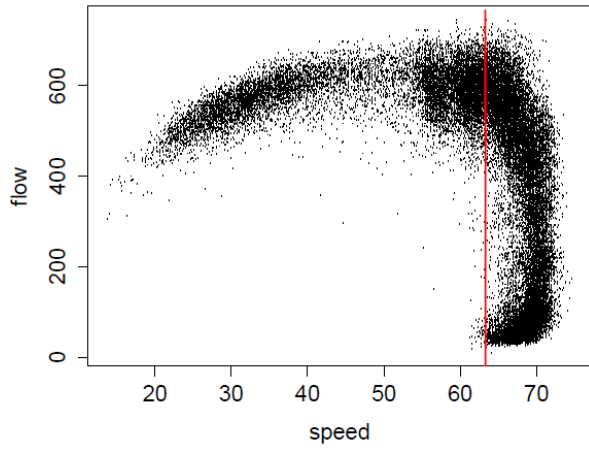
**I-10 E,mile = 4**



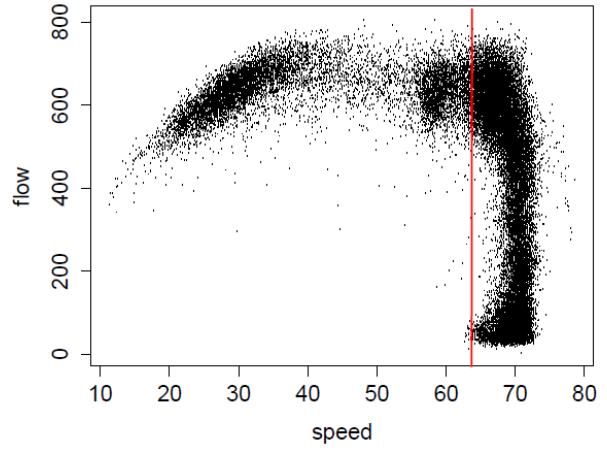
**I-10 E,mile = 4.34**



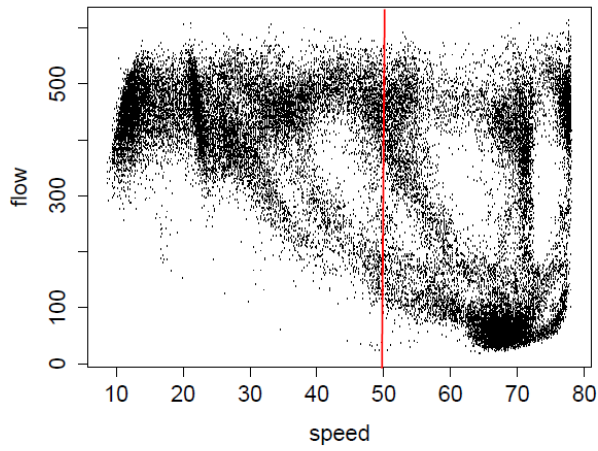
**I-10 E, mile = 4.58**



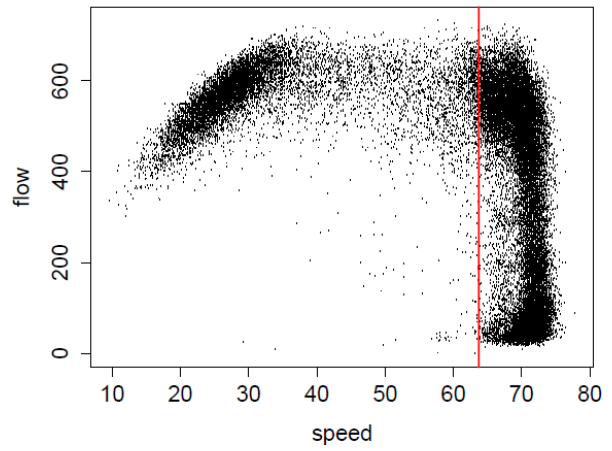
**I-10 E, mile = 5.07**



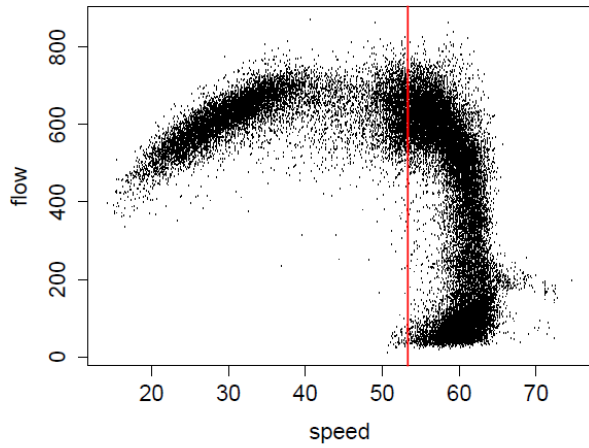
**I-10 E, mile = 5.64**



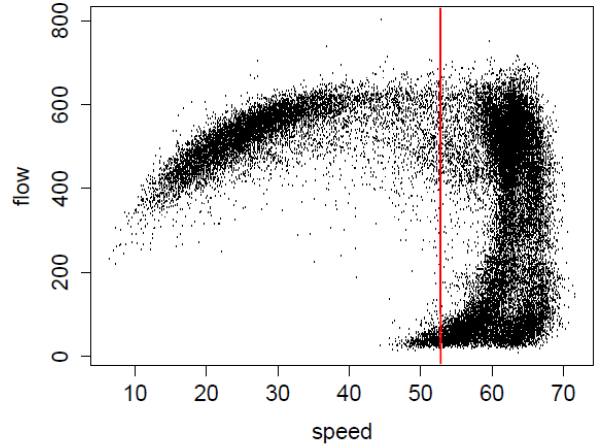
**I-10 E, mile = 5.83**



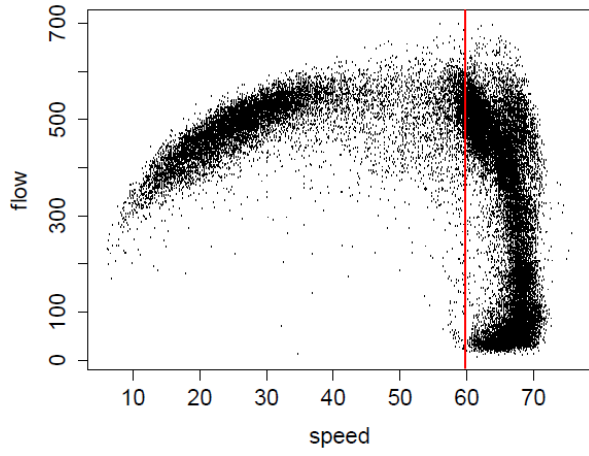
**I-10 E, mile = 6.21**



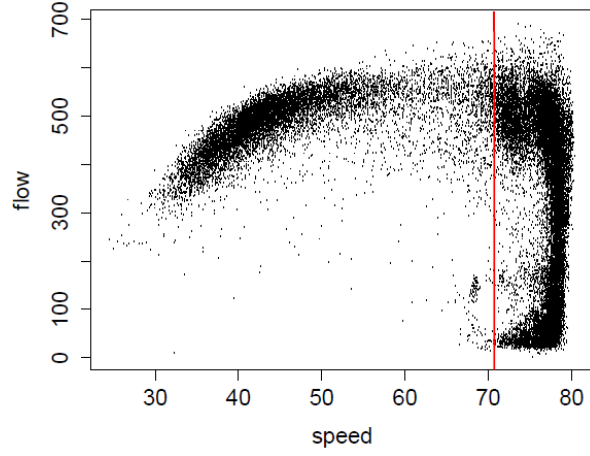
**I-10 E, mile = 6.59**



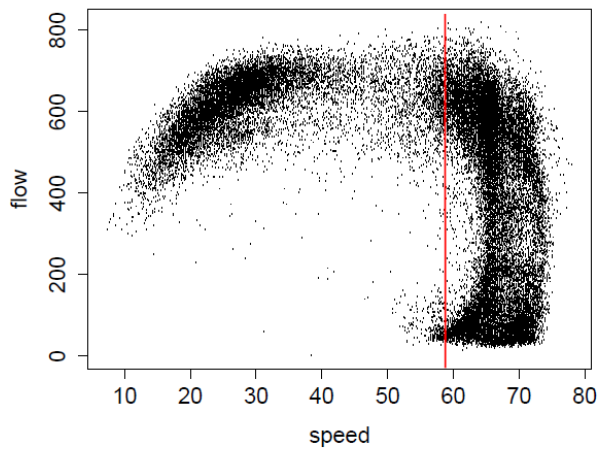
**I-10 E, mile = 6.86**



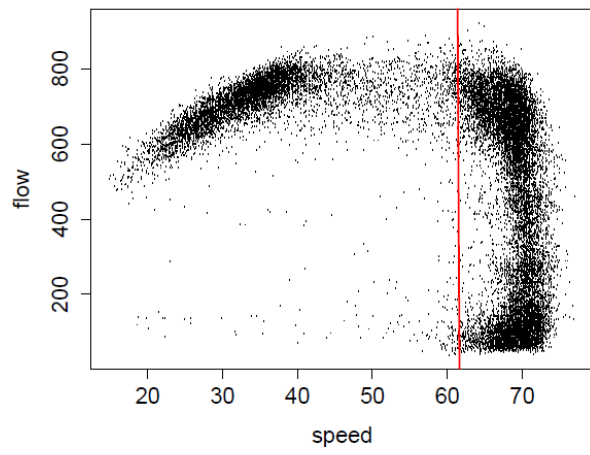
**I-10 E, mile = 7.05**



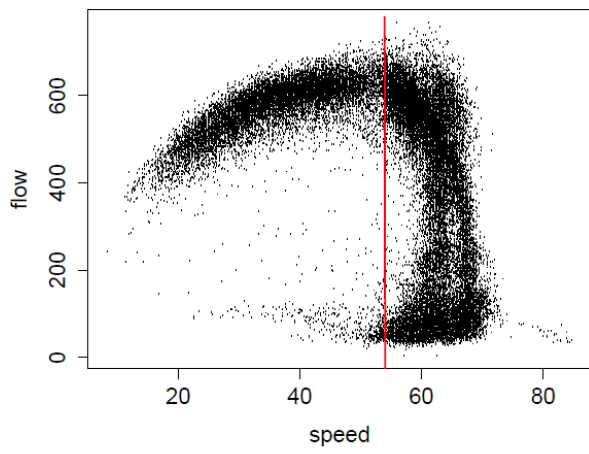
**I-10 E, mile = 7.36**



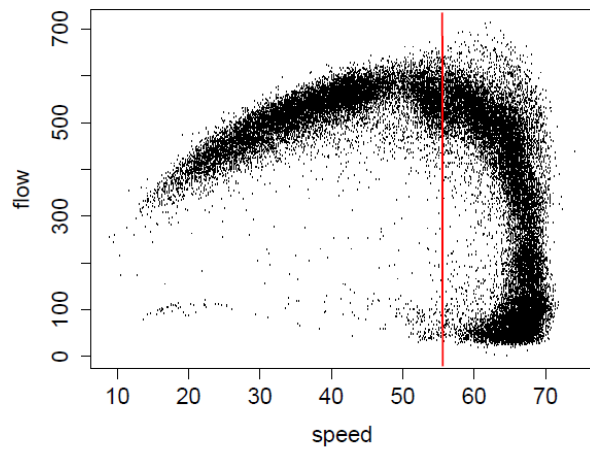
**I-10 E, mile = 7.64**



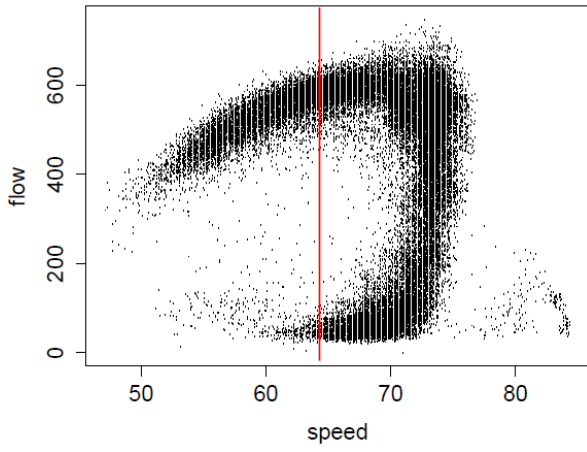
**I-10 E, mile = 8.07**



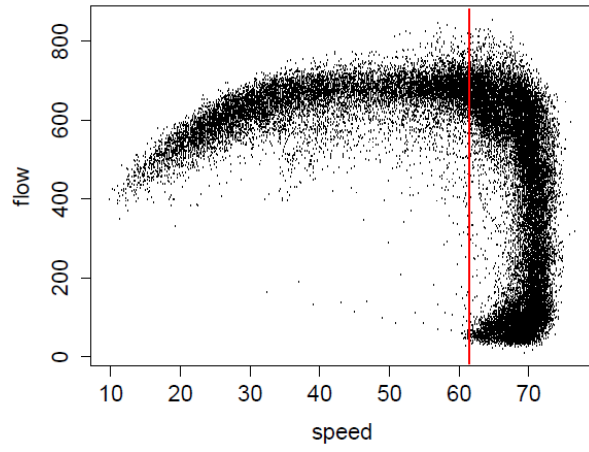
**I-10 E, mile = 8.14**



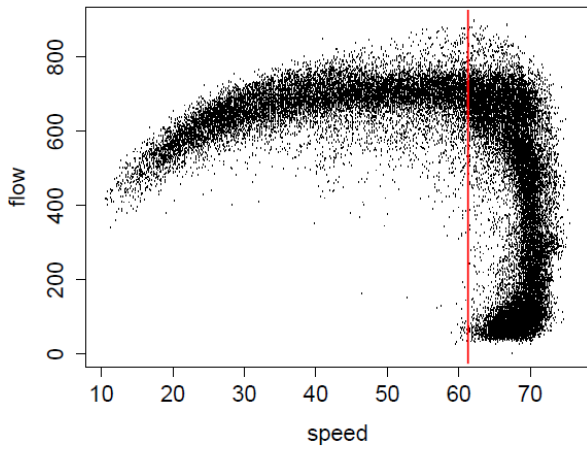
**I-10 E,mile = 8.38**



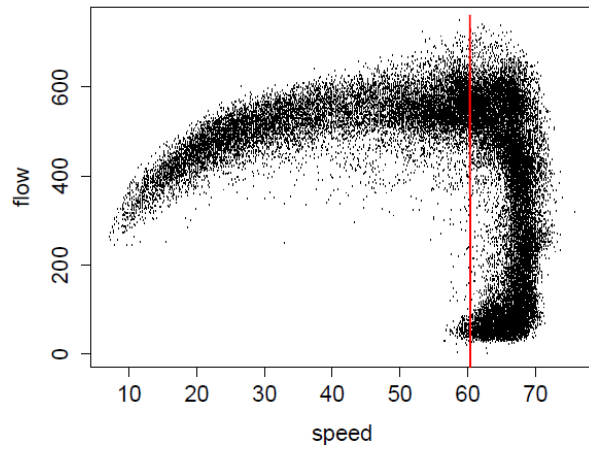
**I-10 E,mile = 9.38**



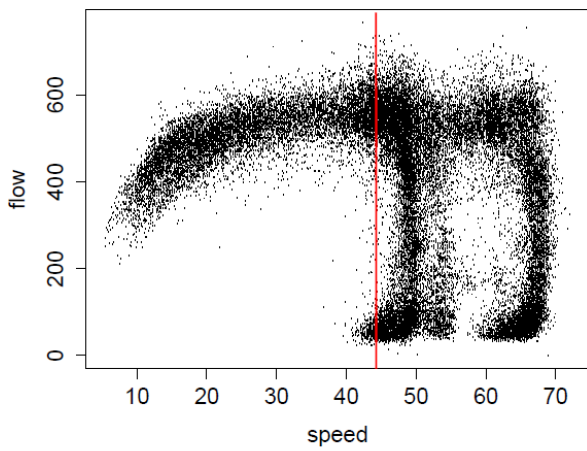
**I-10 E,mile = 10.07**



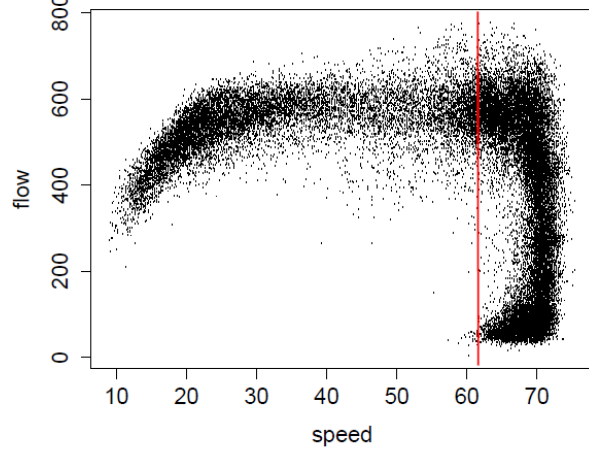
**I-10 E,mile = 10.29**



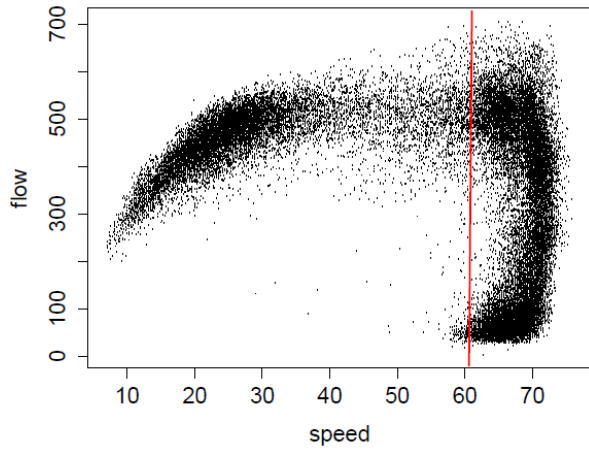
**I-10 E,mile = 10.43**



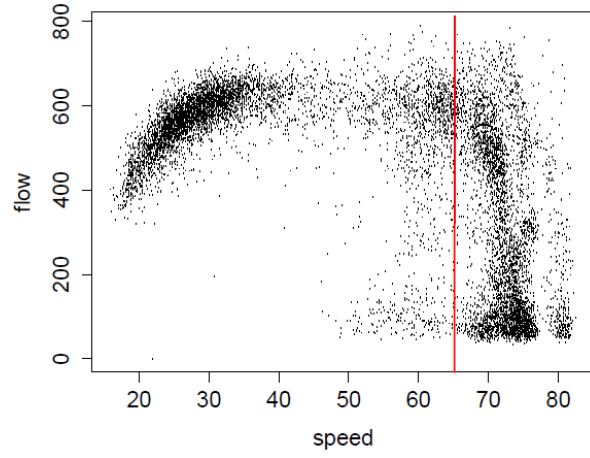
**I-10 E,mile = 10.79**



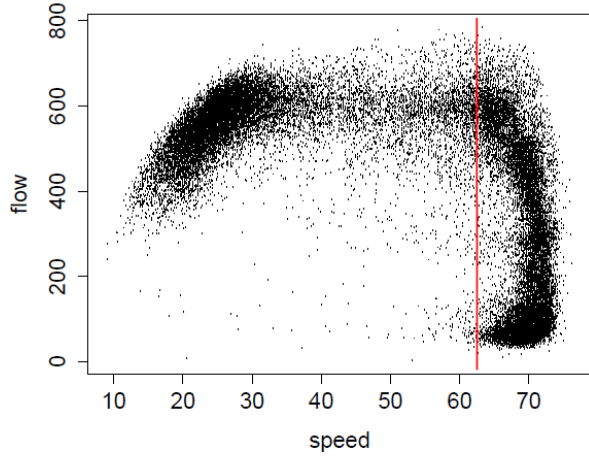
I-10 E, mile = 11.29



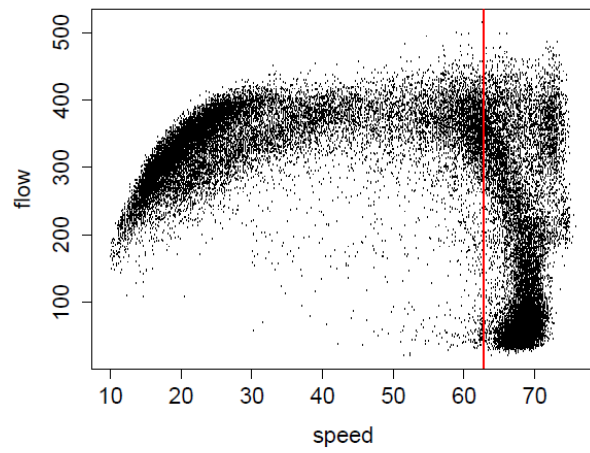
I-10 E, mile = 11.79



I-10 E, mile = 11.96

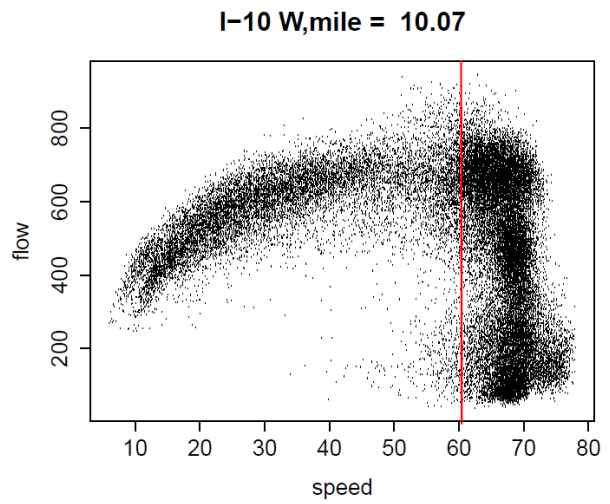
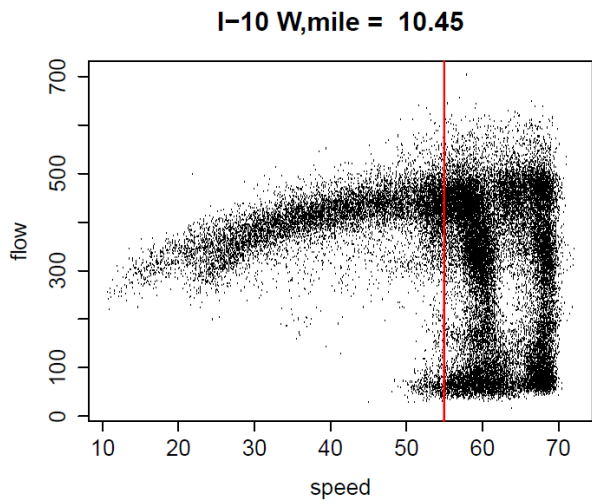
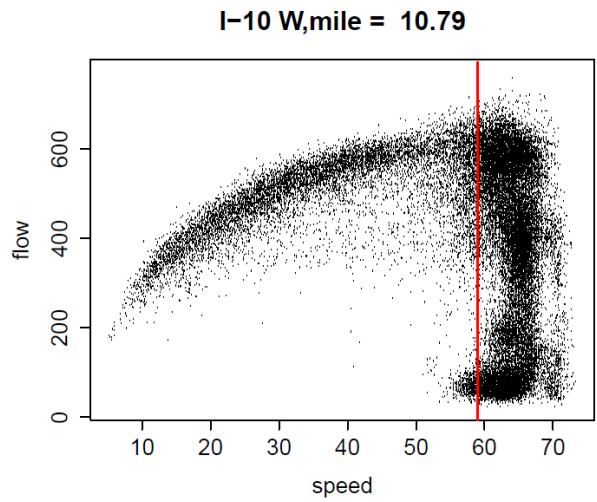
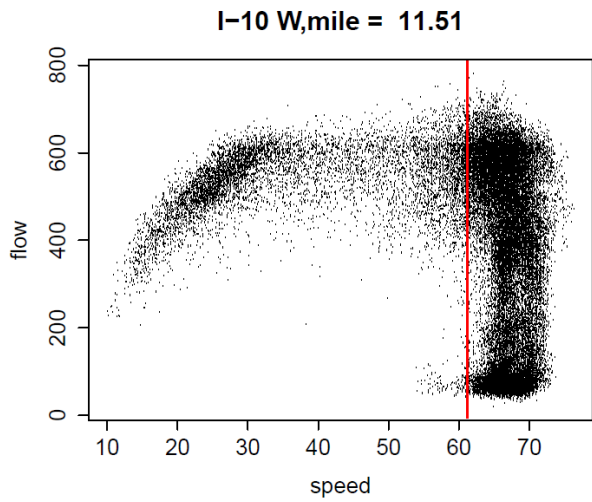
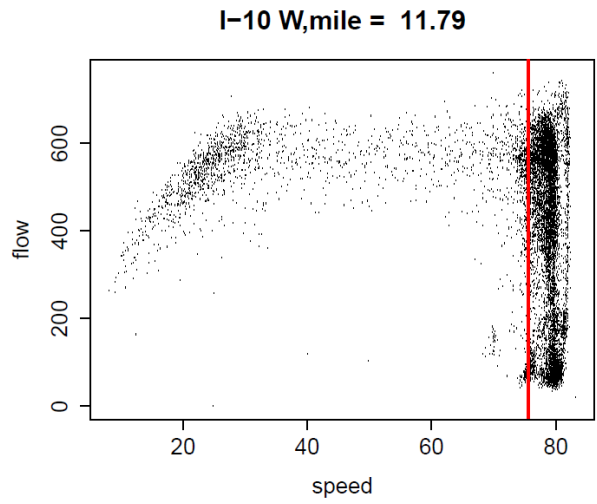
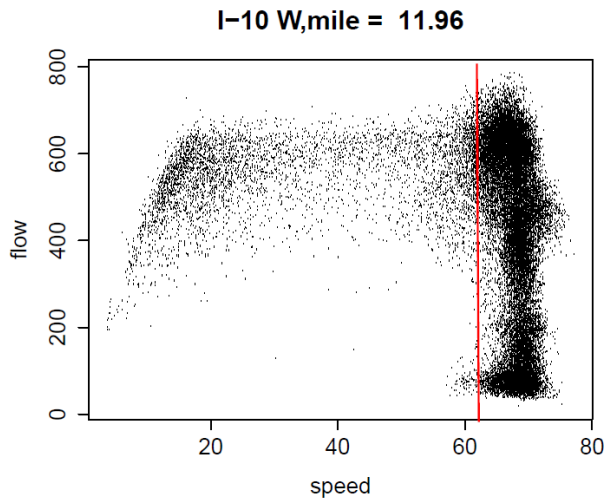


I-10 E, mile = 12.21

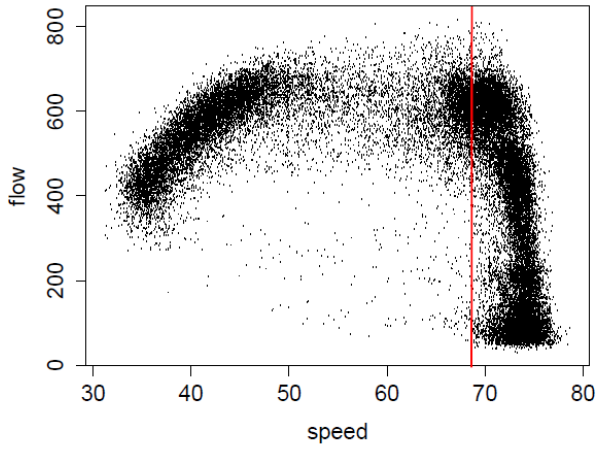




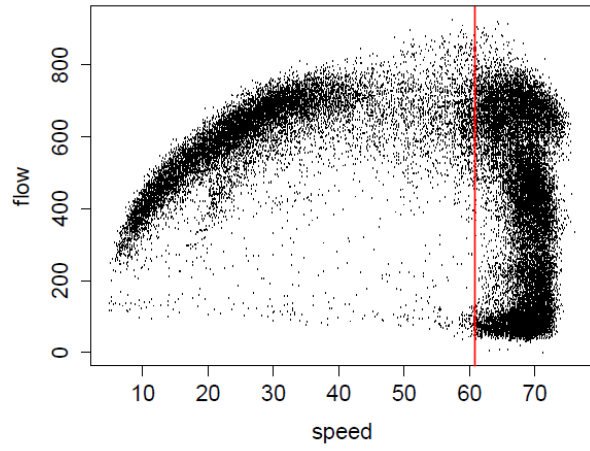
## B. WESTBOUND DIRECTION



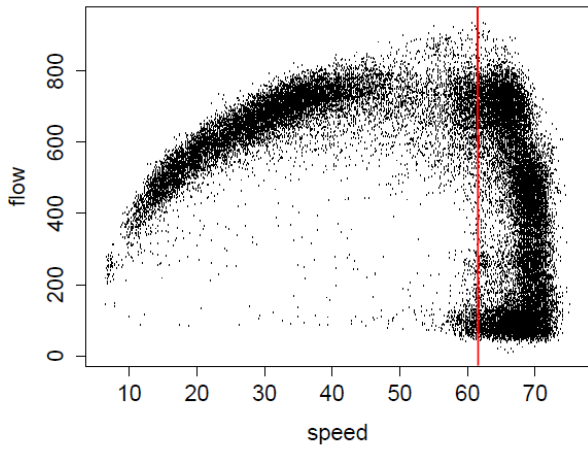
**I-10 W,mile = 9.38**



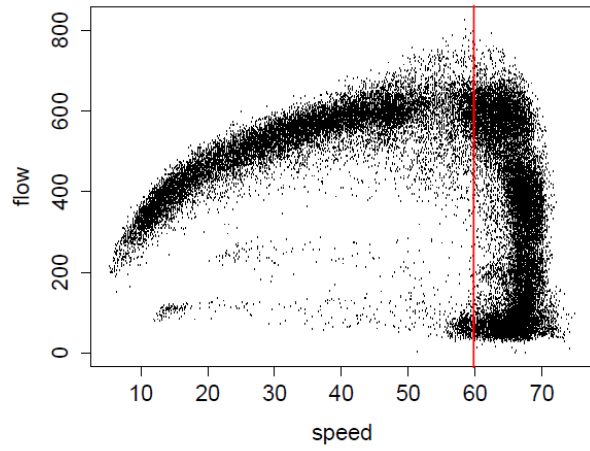
**I-10 W,mile = 9.04**



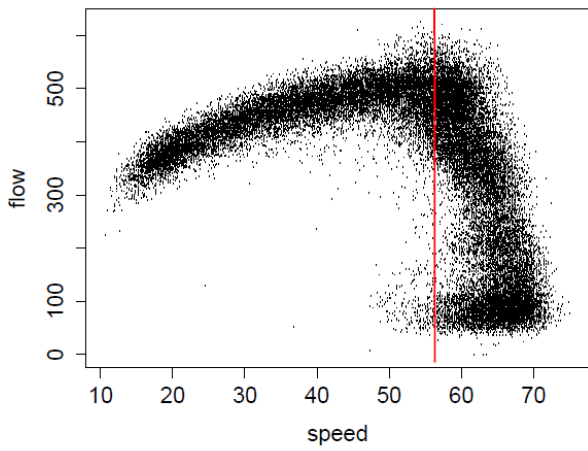
**I-10 W,mile = 8.9**



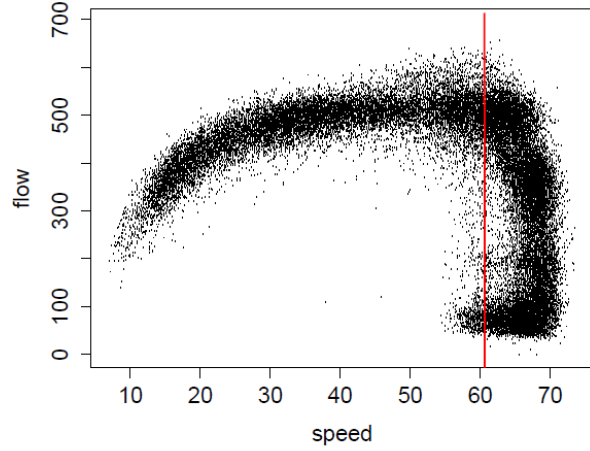
**I-10 W,mile = 8.38**



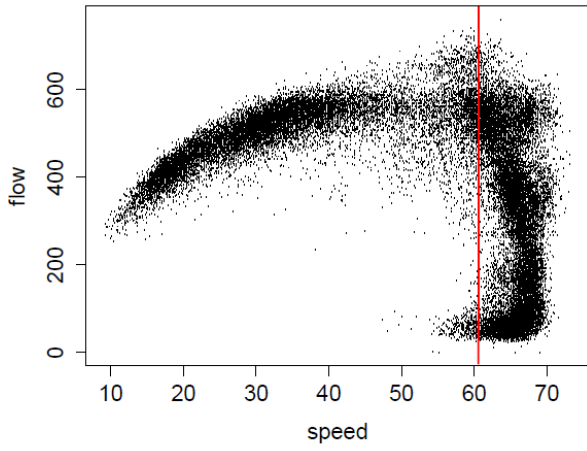
**I-10 W,mile = 8.14**



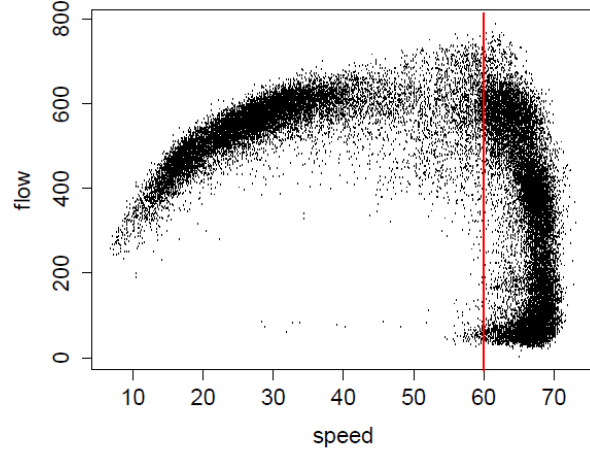
**I-10 W,mile = 7.64**



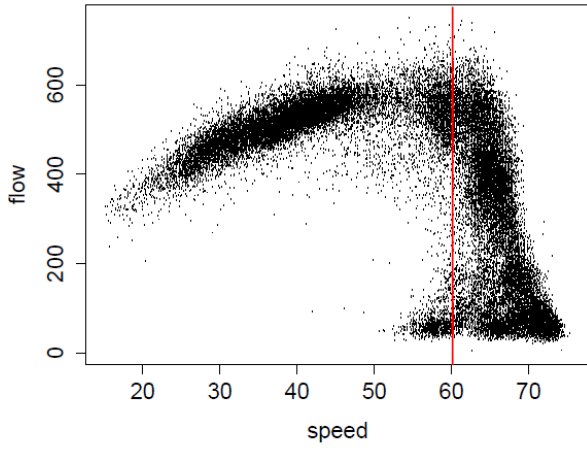
**I-10 W,mile = 7.05**



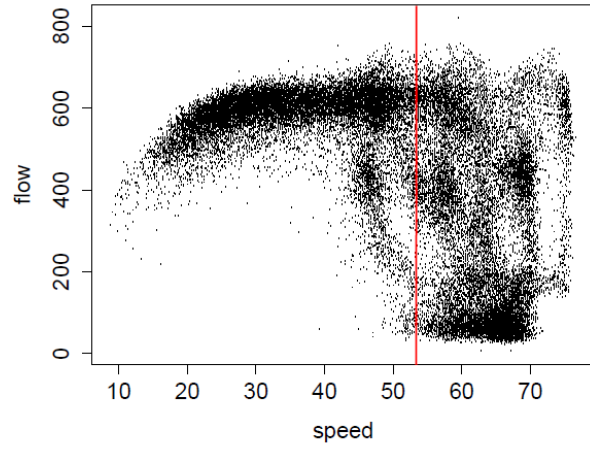
**I-10 W,mile = 6.58**



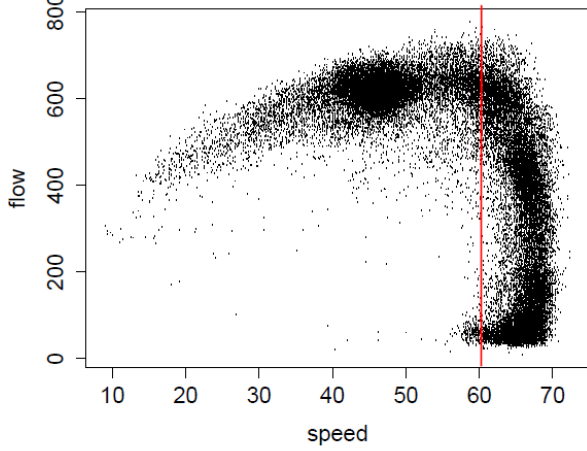
**I-10 W,mile = 6.21**



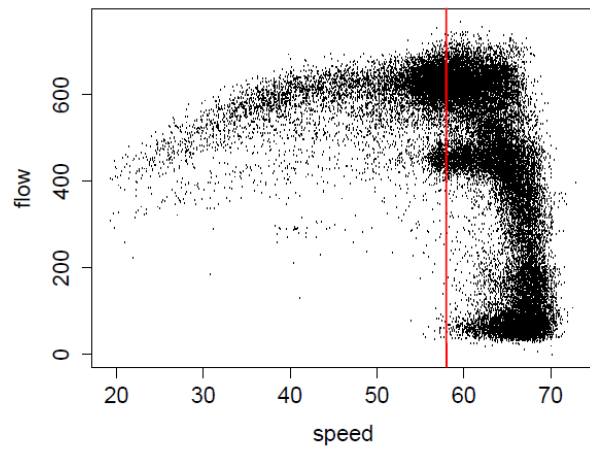
**I-10 W,mile = 5.66**



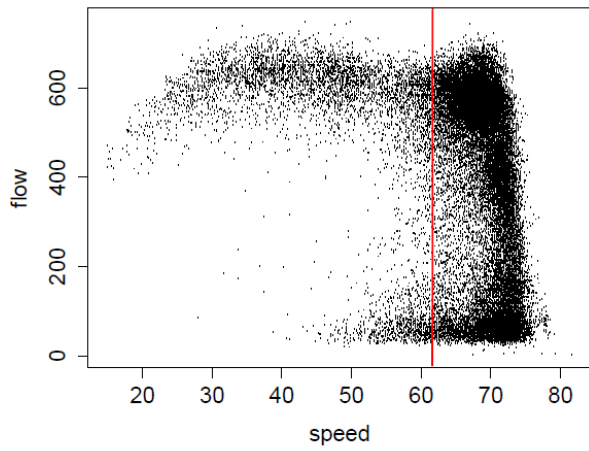
**I-10 W,mile = 5.07**



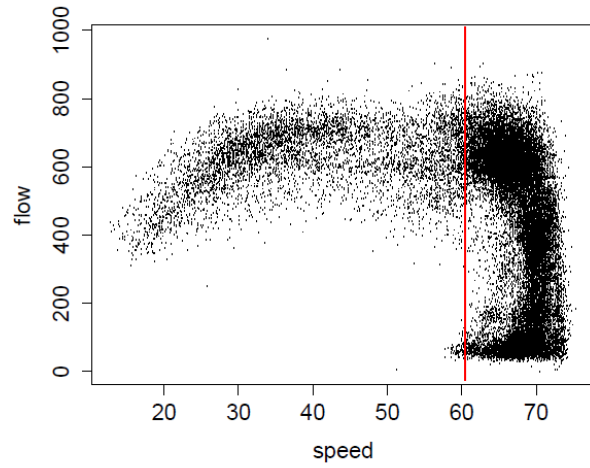
**I-10 W,mile = 4.58**



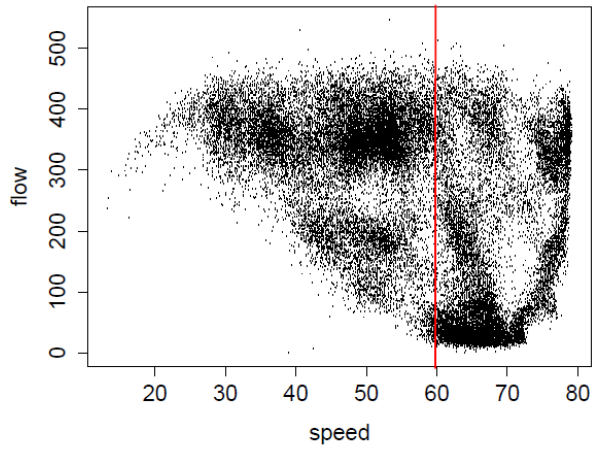
**I-10 W,mile = 4.3**



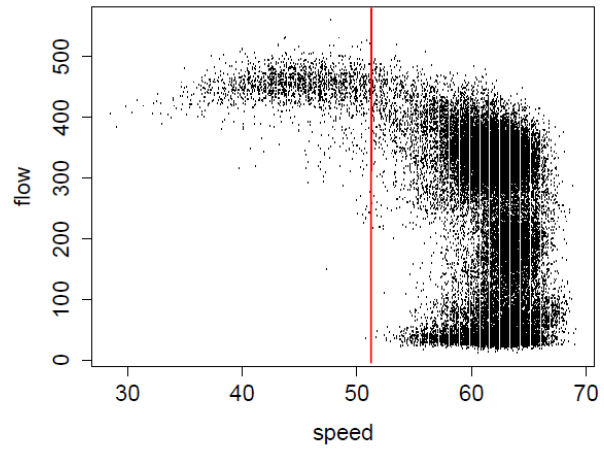
**I-10 W,mile = 4**



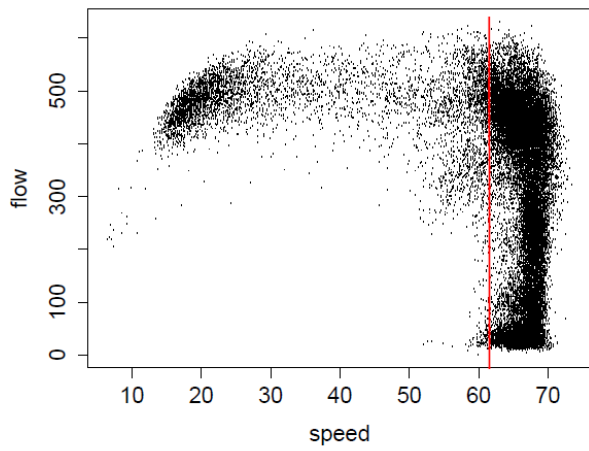
**I-10 W,mile = 3.46**



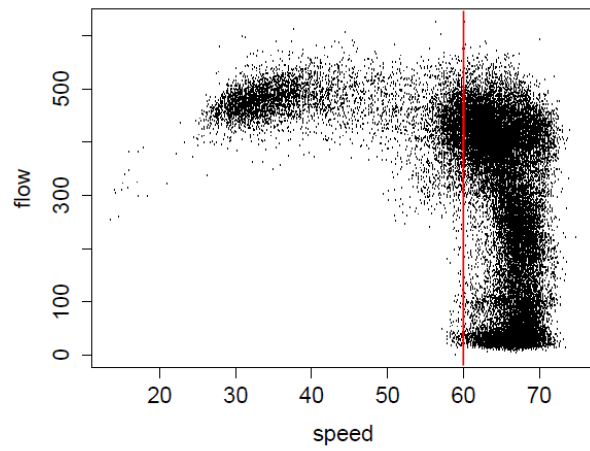
**I-10 W,mile = 3.06**



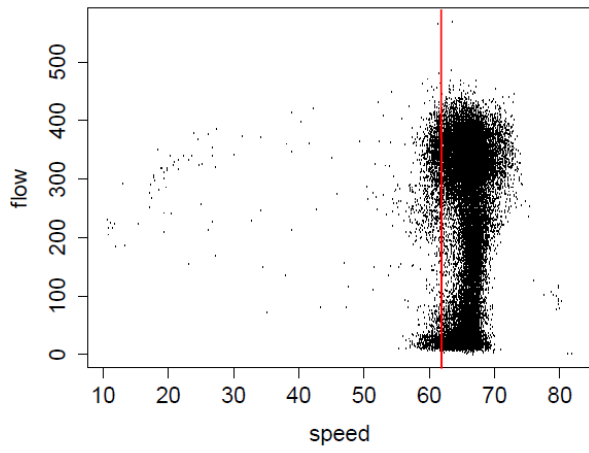
**I-10 W,mile = 1.97**



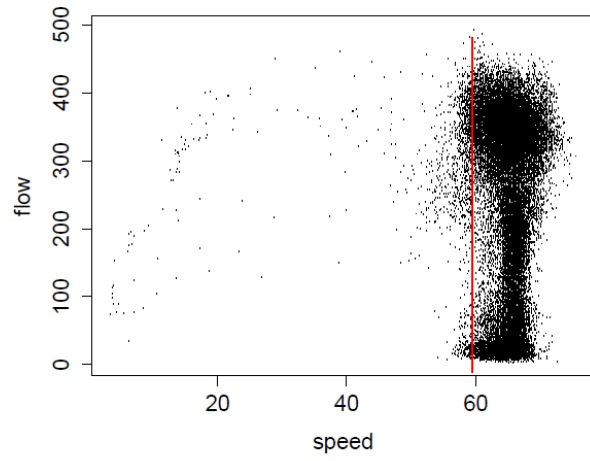
**I-10 W,mile = 1.77**



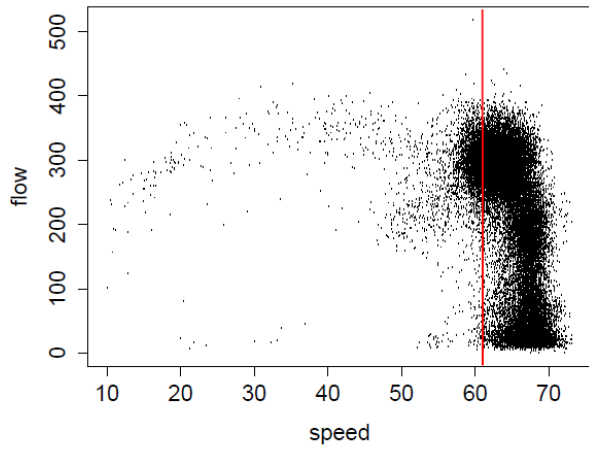
I-10 W,mile = 0.93



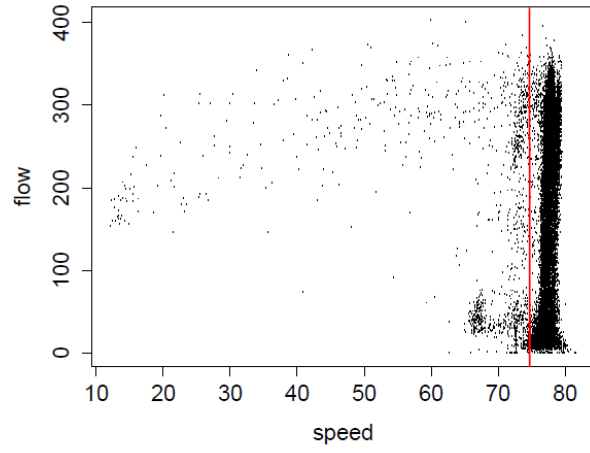
I-10 W,mile = 0.78



I-10 W,mile = 0.47



I-10 W,mile = 0.17



**APPENDIX 4**

**TRAFFIC CONDITION FREQUENCY TABLES FOR DAYTIME PERIODS**

state period

Frequency Percent Row Pct Col Pct	71	72	73	74	75	76	77	78	79	80	Total
0	93 0.01 0.02 2.47	90 0.01 0.02 2.39	91 0.01 0.02 2.40	92 0.01 0.02 2.45	98 0.01 0.03 2.58	99 0.01 0.03 2.61	110 0.01 0.03 2.90	136 0.01 0.04 3.59	134 0.01 0.03 3.53	134 0.01 0.03 3.53	386743 35.52
1	3666 0.34 0.52 97.53	3668 0.34 0.52 97.61	3701 0.34 0.53 97.60	3666 0.34 0.52 97.55	3694 0.34 0.53 97.42	3692 0.34 0.53 97.39	3681 0.34 0.52 97.10	3656 0.34 0.52 96.41	3658 0.34 0.52 96.47	3657 0.34 0.52 96.47	701967 64.48
Total	3759 0.35	3758 0.35	3792 0.35	3758 0.35	3792 0.35	3791 0.35	3791 0.35	3792 0.35	3792 0.35	3791 0.35	1088710 100.00

6 AM - beginning of "day"

(Continued)

state period

Frequency Percent Row Pct Col Pct	81	82	83	84	85	86	87	88	89	90	Total
0	154 0.01 0.04 4.06	205 0.02 0.05 5.41	248 0.02 0.06 6.54	292 0.03 0.08 7.70	361 0.03 0.09 9.52	471 0.04 0.12 12.42	725 0.07 0.19 19.12	1249 0.11 0.32 32.94	1676 0.15 0.43 44.21	1931 0.18 0.50 50.95	386743 35.52
1	3637 0.33 0.52 95.94	3586 0.33 0.51 94.59	3543 0.33 0.50 93.46	3499 0.32 0.50 92.30	3431 0.32 0.49 90.48	3321 0.31 0.47 87.58	3067 0.28 0.44 80.88	2543 0.23 0.36 67.06	2115 0.19 0.30 55.79	1859 0.17 0.26 49.05	701967 64.48
Total	3791 0.35	3791 0.35	3791 0.35	3791 0.35	3792 0.35	3792 0.35	3792 0.35	3792 0.35	3791 0.35	3790 0.35	1088710 100.00

"0" means unstable, or congested condition

morning peak begins. period 90 => 7:30 AM

(Continued)

state period

Frequency Percent Row Pct Col Pct	91	92	93	94	95	96	97	98	99	100	Total
0	2081 0.19 0.54 54.91	2177 0.20 0.56 57.44	2297 0.21 0.59 60.57	2401 0.22 0.62 63.32	2450 0.23 0.63 64.64	2479 0.23 0.64 65.44	2540 0.23 0.66 67.00	2607 0.24 0.67 68.77	2694 0.25 0.70 71.08	2801 0.26 0.72 74.02	386743 35.52
1	1709 0.16 0.24 45.09	1613 0.15 0.23 42.56	1495 0.14 0.21 39.43	1391 0.13 0.20 36.68	1340 0.12 0.19 35.36	1309 0.12 0.19 34.56	1251 0.11 0.18 33.00	1184 0.11 0.17 31.23	1096 0.10 0.16 28.92	983 0.09 0.14 25.98	701967 64.48
Total	3790 0.35	3790 0.35	3792 0.35	3792 0.35	3790 0.35	3788 0.35	3791 0.35	3791 0.35	3790 0.35	3784 0.35	1088710 100.00

peak criterion: over 50% unstable

(Continued)

state period

Frequency											
Percent											
Row Pct											
Col Pct	101	102	103	104	105	106	107	108	109	110	Total
0	2865	2889	2911	2895	2856	2794	2749	2657	2537	2421	386743
	0.26	0.27	0.27	0.27	0.26	0.26	0.25	0.24	0.23	0.22	35.52
	0.74	0.75	0.75	0.75	0.74	0.72	0.71	0.69	0.66	0.63	
	75.71	76.35	76.93	76.39	75.36	73.72	72.53	70.20	67.03	63.96	
1	919	895	873	895	934	996	1041	1128	1248	1364	701967
	0.08	0.08	0.08	0.08	0.09	0.09	0.10	0.10	0.11	0.13	64.48
	0.13	0.13	0.12	0.13	0.13	0.14	0.15	0.16	0.18	0.19	
	24.29	23.65	23.07	23.61	24.64	26.28	27.47	29.80	32.97	36.04	
Total	3784	3784	3784	3790	3790	3790	3790	3785	3785	3785	1088710
	0.35	0.35	0.35	0.35				0.35	0.35	0.35	100.00

(Continued)

state period

Frequency											
Percent											
Row Pct											
Col Pct	111	112	113	114	115	116	117	118	119	120	Total
0	2295	2220	2147	2081	2021	1996	1957	1905	1795	1622	386743
	0.21	0.20	0.20	0.19	0.19	0.18	0.18	0.17	0.16	0.15	35.52
	0.59	0.57	0.56	0.54	0.52	0.52	0.51	0.49	0.46	0.42	
	60.63	58.65	56.72	54.98	53.52	52.86	51.83	50.45	47.52	42.96	
1	1490	1565	1638	1704	1755	1780	1819	1871	1982	2154	701967
	0.14	0.14	0.15	0.16	0.16	0.16	0.17	0.17	0.18	0.20	64.48
	0.21	0.22	0.23	0.24	0.25	0.25	0.26	0.27	0.28	0.31	
	39.37	41.35	43.28	45.02	46.48	47.14	48.17	49.55	52.48	57.04	
Total	3785	3785	3785	3785	3776	3776	3776	3776	3777	3776	1088710
	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	100.00

(Continued)

state period

Frequency											
Percent											
Row Pct											
Col Pct	121	122	123	124	125	126	127	128	129	130	Total
0	1536	1433	1341	1249	1185	1202	1191	1177	1183	1172	386743
	0.14	0.13	0.12	0.11	0.11	0.11	0.11	0.11	0.11	0.11	35.52
	0.40	0.37	0.35	0.32	0.31	0.31	0.31	0.30	0.31	0.30	
	40.68	37.95	35.77	33.32	31.63	32.12	31.84	31.48	31.64	31.07	
1	2240	2343	2408	2500	2562	2540	2549	2562	2556	2600	701967
	0.21	0.22	0.22	0.23	0.24	0.23	0.23	0.24	0.23	0.24	64.48
	0.32	0.33	0.34	0.36	0.36	0.36	0.36	0.36	0.36	0.37	
	59.32	62.05	64.23	66.68	68.37	67.88	68.16	68.52	68.36	68.93	
Total	3776	3776	3749	3749	3747	3742	3740	3739	3739	3772	1088710
	0.35	0.35	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.35	100.00

(Continued)

morning peak ends.  
period 118 -> 9:50 AM



state period

Frequency Percent Row Pot Col Pot	131	132	133	134	135	136	137	138	139	140	Total
0	1137 0.10 0.29 30.08	1118 0.10 0.29 29.79	1094 0.10 0.28 29.16	1109 0.10 0.29 29.55	1155 0.11 0.30 30.60	1168 0.11 0.30 31.20	1173 0.11 0.30 31.07	1189 0.11 0.31 31.52	1238 0.11 0.32 32.85	1245 0.11 0.32 33.03	386743 35.52
1	2643 0.24 0.38 69.92	2635 0.24 0.38 70.21	2658 0.24 0.38 70.84	2644 0.24 0.38 70.45	2620 0.24 0.37 69.40	2575 0.24 0.37 68.80	2602 0.24 0.37 68.93	2583 0.24 0.37 68.48	2531 0.23 0.36 67.15	2524 0.23 0.36 66.97	701967 64.48
Total	3780 0.35	3753 0.34	3752 0.34	3753 0.34	3775 0.35	3743 0.34	3775 0.35	3772 0.35	3769 0.35	3769 0.35	1088710 100.00

(Continued)

state period

Frequency Percent Row Pot Col Pot	141	142	143	144	145	146	147	148	149	150	Total
0	1259 0.12 0.33 33.40	1285 0.12 0.33 34.19	1288 0.12 0.33 34.27	1272 0.12 0.33 33.85	1269 0.12 0.33 33.77	1288 0.12 0.33 34.27	1334 0.12 0.34 35.50	1364 0.13 0.35 36.31	1425 0.13 0.37 37.93	1466 0.13 0.38 39.32	386743 35.52
1	2510 0.23 0.36 66.60	2473 0.23 0.35 65.81	2470 0.23 0.35 65.73	2486 0.23 0.35 66.15	2489 0.23 0.35 66.23	2470 0.23 0.35 65.73	2424 0.22 0.35 64.50	2393 0.22 0.34 63.69	2332 0.21 0.33 62.07	2262 0.21 0.32 60.68	701967 64.48
Total	3769 0.35	3758 0.35	3758 0.35	3758 0.35	3758 0.35	3758 0.35	3758 0.35	3757 0.35	3757 0.35	3728 0.34	1088710 100.00

(Continued)

state period

Frequency Percent Row Pot Col Pot	151	152	153	154	155	156	157	158	159	160	Total
0	1522 0.14 0.39 40.51	1555 0.14 0.40 41.39	1615 0.15 0.42 42.99	1624 0.15 0.42 43.32	1615 0.15 0.42 43.10	1617 0.15 0.42 43.15	1640 0.15 0.42 43.77	1635 0.15 0.42 43.89	1638 0.15 0.42 43.66	1709 0.16 0.44 45.55	386743 35.52
1	2235 0.21 0.32 59.49	2202 0.20 0.31 58.61	2142 0.20 0.31 57.01	2125 0.20 0.30 56.68	2132 0.20 0.30 56.90	2130 0.20 0.30 56.85	2107 0.19 0.30 56.23	2090 0.19 0.30 56.11	2114 0.19 0.30 56.34	2043 0.19 0.29 54.45	701967 64.48
Total	3757 0.35	3757 0.35	3757 0.35	3749 0.34	3747 0.34	3747 0.34	3747 0.34	3725 0.34	3752 0.34	3752 0.34	1088710 100.00

(Continued)

state		period												
Frequency	Percent	Row Pct	Col Pct	161	162	163	164	165	166	167	168	169	170	Total
0	1804	1879	1913	1965	1985	2047	2089	2135	2110	2128			386743	
	0.17	0.17	0.18	0.18	0.18	0.19	0.19	0.20	0.19	0.20			35.52	
	0.47	0.49	0.49	0.51	0.51	0.53	0.54	0.55	0.55	0.55				
	48.03	50.03	50.86	52.25	53.19	54.57	55.51	56.74	56.07	56.55				
1	1952	1877	1848	1796	1747	1704	1674	1628	1653	1635			701967	
	0.18	0.17	0.17	0.16	0.16	0.16	0.15	0.15	0.15	0.15			64.48	
	0.28	0.27	0.26	0.26	0.25	0.24	0.24	0.23	0.24	0.23				
	51.97	49.97	49.14	47.75	46.81	45.43	44.49	43.26	43.93	43.45				
Total	3756	3756	3761	3761	3732	3751	3763	3763	3763	3763	3763	3763	1088710	
	0.34	0.34	0.35	0.35	0.34	0.34	0.35	0.35	0.35	0.35	0.35	0.35	100.00	

(Continued)

afternoon peak begins.  
period 162 -> 1:25 PM

state		period												
Frequency	Percent	Row Pct	Col Pct	171	172	173	174	175	176	177	178	179	180	Total
0	2201	2336	2514	2650	2757	2846	2923	2993	2999	2999	3013			386743
	0.20	0.21	0.23	0.24	0.25	0.26	0.27	0.27	0.27	0.28	0.28			35.52
	0.57	0.60	0.65	0.69	0.71	0.74	0.76	0.77	0.78	0.78	0.78			
	58.57	62.16	66.81	70.42	73.36	75.73	77.78	79.58	79.74	80.11				
1	1557	1422	1249	1113	1001	912	835	768	762	748			701967	
	0.14	0.13	0.11	0.10	0.09	0.08	0.08	0.07	0.07	0.07			64.48	
	0.22	0.20	0.18	0.16	0.14	0.13	0.12	0.11	0.11	0.11				
	41.43	37.84	33.19	29.58	26.64	24.27	22.22	20.42	20.26	19.89				
Total	3758	3758	3763	3763	3758	3758	3758	3761	3761	3761	3761	3761	1088710	
	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	100.00	

(Continued)

state		period												
Frequency	Percent	Row Pct	Col Pct	181	182	183	184	185	186	187	188	189	190	Total
0	3009	3012	3060	3142	3252	3309	3320	3340	3331	3358			386743	
	0.28	0.28	0.28	0.29	0.30	0.30	0.30	0.31	0.31	0.31			35.52	
	0.78	0.78	0.79	0.81	0.84	0.86	0.86	0.86	0.86	0.87				
	80.01	80.09	81.36	84.28	86.47	87.98	88.27	88.81	89.09	89.52				
1	752	749	701	586	509	452	441	421	408	393			701967	
	0.07	0.07	0.06	0.05	0.05	0.04	0.04	0.04	0.04	0.04			64.48	
	0.11	0.11	0.10	0.08	0.07	0.06	0.06	0.06	0.06	0.06				
	19.99	19.91	18.64	15.72	13.53	12.02	11.73	11.19	10.91	10.48				
Total	3761	3761	3761	3728	3761	3761	3761	3761	3761	3739	3751	3751	1088710	
	0.35	0.35	0.35	0.34	0.35	0.35	0.35	0.35	0.35	0.34	0.34	0.34	100.00	

(Continued)

state period

Frequency Percent Row Pct Col Pct	→										Total
	191	192	193	194	195	196	197	198	199	200	
0	3360	3347	3324	3297	3313	3327	3333	3332	3316	3288	386743
	0.31	0.31	0.31	0.30	0.30	0.31	0.31	0.31	0.30	0.30	35.52
	0.87	0.87	0.86	0.85	0.86	0.86	0.86	0.86	0.86	0.85	
	89.67	89.23	88.90	88.25	88.28	88.44	88.57	88.55	88.12	87.38	
1	387	404	415	439	440	435	430	431	447	475	701967
	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	64.48
	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.07	
	10.33	10.77	11.10	11.75	11.72	11.56	11.43	11.45	11.88	12.62	
Total	3747	3751	3739	3736	3753	3762	3763	3763	3763	3763	1088710
	0.34	0.34	0.34	0.34	0.34	0.35	0.35	0.35	0.35	0.35	100.00

(Continued)

state period

Frequency Percent Row Pct Col Pct	→										Total
	201	202	203	204	205	206	207	208	209	210	
0	3256	3246	3224	3201	3154	3155	3177	3217	3239	3269	386743
	0.30	0.30	0.30	0.29	0.29	0.29	0.29	0.30	0.30	0.30	35.52
	0.84	0.84	0.83	0.83	0.82	0.82	0.82	0.83	0.84	0.85	
	86.76	86.49	85.90	85.04	83.79	83.60	84.20	85.26	85.82	86.64	
1	497	507	529	563	610	619	596	556	535	504	701967
	0.05	0.05	0.05	0.05	0.06	0.06	0.05	0.05	0.05	0.05	64.48
	0.07	0.07	0.08	0.08	0.09	0.09	0.08	0.08	0.08	0.07	
	13.24	13.51	14.10	14.96	16.21	16.40	15.80	14.74	14.18	13.36	
Total	3753	3753	3753	3764	3764	3774	3773	3773	3774	3773	1088710
	0.34	0.34	0.34	0.35	0.35	0.35	0.35	0.35	0.35	0.35	100.00


(Continued)

state period

Frequency Percent Row Pct Col Pct	→										Total
	211	212	213	214	215	216	217	218	219	220	
0	3282	3285	3306	3293	3268	3253	3224	3180	3155	3165	386743
	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.29	0.29	0.29	35.52
	0.85	0.85	0.85	0.85	0.85	0.84	0.83	0.82	0.82	0.82	
	86.99	87.07	87.62	87.28	86.62	86.22	85.45	84.28	83.62	83.89	
1	491	488	467	480	505	520	549	593	618	608	701967
	0.05	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.06	64.48
	0.07	0.07	0.07	0.07	0.07	0.07	0.08	0.08	0.09	0.09	
	13.01	12.93	12.38	12.72	13.38	13.78	14.55	15.72	16.38	16.11	
Total	3773	3773	3773	3773	3773	3773	3773	3773	3773	3773	1088710
	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	100.00

(Continued)


state period




Frequency	Percent	Row Pct	Col Pct	221	222	223	224	225	226	227	228	229	230	Total
0	3173	3163	3138	3115	3053	2987	2905	2796	2668	2545				386743
	0.29	0.29	0.29	0.29	0.28	0.27	0.27	0.26	0.25	0.23				35.52
	0.82	0.82	0.81	0.81	0.79	0.77	0.75	0.72	0.69	0.66				
	84.10	83.83	83.19	82.60	80.90	79.15	76.97	74.09	70.69	67.44				
1	600	610	634	656	721	787	869	978	1106	1229				701967
	0.06	0.06	0.06	0.06	0.07	0.07	0.08	0.09	0.10	0.11				64.48
	0.09	0.09	0.09	0.09	0.10	0.11	0.12	0.14	0.16	0.18				
	15.90	16.17	16.81	17.40	19.10	20.85	23.03	25.91	29.31	32.56				
Total	3773	3773	3772	3771	3774	3774	3774	3774	3774	3774	3774	3774	3774	1088710
	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	100.00

(Continued)


state period



afternoon peak ends. period 237 -> 7:45 PM



8 PM, end of "day"



Frequency	Percent	Row Pct	Col Pct	231	232	233	234	235	236	237	238	239	240	Total
0	2459	2375	2317	2217	2132	2046	1936	1784	1618	1427				386743
	0.23	0.22	0.21	0.20	0.20	0.19	0.18	0.16	0.15	0.13				35.52
	0.64	0.61	0.60	0.57	0.55	0.53	0.50	0.46	0.42	0.37				
	65.10	62.88	61.77	59.10	57.27	54.55	51.61	47.56	43.14	38.04				
1	1318	1402	1434	1534	1591	1705	1815	1967	2133	2324				701967
	0.12	0.13	0.13	0.14	0.15	0.16	0.17	0.18	0.20	0.21				64.48
	0.19	0.20	0.20	0.22	0.23	0.24	0.26	0.28	0.30	0.33				
	34.90	37.12	38.23	40.90	42.73	45.45	48.39	52.44	56.86	61.96				
Total	3777	3777	3751	3751	3723	3751	3751	3751	3751	3751	3751	3751	3751	1088710
	0.35	0.35	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	100.00

(Continued)

## BIBLIOGRAPHY

1. Asuero, A., Sayago, A., and González, A. The Correlation Coefficient: An Overview. *Critical Reviews in Analytical Chemistry*. 36:1. Seville, Spain, 2006, pp. 41-59.
2. Bartlett, R. Linear Modeling of Pearson's Product Moment Correlation Coefficient: An Application of Fisher's Z Transformation. *Journal of the Royal Statistical Society. Series D (The Statistician)*, Vol. 42, No. 1, 1993, pp. 45-53.
3. Chandra, S., and Al-Deel H. Cross-Correlation Analysis and Multivariate Prediction of Spatial Time Series of Freeway Traffic Speeds. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2061, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 64-76.
4. Duthie, J., A. Unnikrishnan, and S. T. Waller. Influence of Demand Uncertainty and Correlations on Traffic Predictions and Decisions. *Computer-Aided Civil and Infrastructure Engineering*, Vol. 26, No. 1, 2011, pp. 16-29.
5. Eom, J. K., M.S. Park, H. Tae-Young, and L.F. Huntsinger. Improving the Prediction of Annual Average Daily Traffic for Nonfreeway Facilities by Applying a Spatial Statistical Method. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1968, Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 20-29.
6. Fisher, R. The General Sampling Distribution of the Multiple Correlation Coefficient. *Proceedings of the Royal Statistical Society*. Vol. 121, 1928, pp. 654-673.
7. Frejinger, E., and Bierlaire, M. Capturing Correlation with Subnetworks in Route Choice Models. *Transportation Research Part :Methodological*. Vol. 41, No. 1, 2007, pp. 363-378.
8. Gajewski, B.J, and L.R. Rilett. Estimating Link Travel Time Correlation: An Application of Bayesian Smoothing Splines. *Journal of Transportation and Statistics*, 7(2/3), 2004, pp. 53-70.
9. Gao, S., and I. Chabini. Optimal Routing Policy Problems in Stochastic Time-Dependent Networks. *Transportation Research. Part B: Methodological*, Vol. 40, No. 2, 2006, pp. 93-122.
10. Goel, P. K., M. R. McCord, and C. Park. Exploiting Correlations between Link Flows to Improve AADT Estimation on Coverage Count Segments: Methodology and Numerical Study. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1917, Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 100-107.

11. Jia, Z., C. Chen, B. Coifman, and P. Varaiya. The PeMS Algorithms for Accurate, Real-Time Estimates of g-Factors and Speeds from Single-Loop Detectors. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC* , 2001, pp. 536-541.
12. Min,W., and L. Wynter. Real-Time Road Traffic Prediction with Spatio-Temporal Correlations. *Transportation Research.Part C, Emerging Technologies*, Vol. 19, No. 4, 2011, pp. 606-616.
13. Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. Applied Linear Statistical Models. Irwin, Inc., 3rd ed., 1996.
14. Parent,O., and J. P. LeSage. A Spatial Dynamic Panel Model with Random Effects Applied to Commuting Times. *Transportation Research.Part E, Logistics and Transportation Review*, Vol. 44, No. 5, 2010, pp. 633.
15. Performance Measurement System (PeMS). Station Inventory Report. I-10 E, District 7. From website. Accessed November 2011.  
[http://pems.dot.ca.gov/?report\\_form=1&dnode=Freeway&content=elv&tab=stations&fwy=10&dir=E&\\_time\\_id=1268352000&\\_mm=3&\\_dd=12&\\_yy=2010&eqpo=&st\\_cd=on&st\\_ff=on&st\\_hv=on&st\\_ml=on&st\\_fr=on&st\\_or=on](http://pems.dot.ca.gov/?report_form=1&dnode=Freeway&content=elv&tab=stations&fwy=10&dir=E&_time_id=1268352000&_mm=3&_dd=12&_yy=2010&eqpo=&st_cd=on&st_ff=on&st_hv=on&st_ml=on&st_fr=on&st_or=on)
16. \_\_\_\_\_. Station Inventory Report. I-10 W, District 7. From website. Accessed November 2011. [http://pems.dot.ca.gov/?report\\_form=1&dnode=Freeway&content=elv&tab=stations&fwy=10&dir=W&\\_time\\_id=1268352000&\\_mm=3&\\_dd=12&\\_yy=2010&eqpo=&st\\_cd=on&st\\_ff=on&st\\_hv=on&st\\_ml=on&st\\_fr=on&st\\_or=on](http://pems.dot.ca.gov/?report_form=1&dnode=Freeway&content=elv&tab=stations&fwy=10&dir=W&_time_id=1268352000&_mm=3&_dd=12&_yy=2010&eqpo=&st_cd=on&st_ff=on&st_hv=on&st_ml=on&st_fr=on&st_or=on)
17. Rice,J., and E. van Zwet. A Simple and Effective Method for Predicting Travel Times on Freeways. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 5, No. 3, 2004, pp. 200 -207.
18. Samaranayake, S., Blandin, S., and Bayen, A. Learning the Dependency Structure of Highway Networks for Traffic Forecast. *Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*. Orlando, FL, December 12- 15, 2011
19. Song, X., H. Shao, and G. Wang. Modeling Correlation between Origin-Destination Traffic Demands in Stochastic Transportation Networks. *Proceedings of the 2009 International Joint Conference on Computational Sciences and Optimization, CSO 2009*, Vol. 2 , 2009, pp. 147-149.
20. Transportation Research Board (TRB). Critical Issues in Transportation, 2009 Update. From website. *TRB General Resources*. 2009. Accessed June 2012. <http://onlinepubs.trb.org/Onlinepubs/general/CriticalIssues09.pdf>

21. Tam, M. L., and W. H. K Lam. Short-term Travel Time Prediction for Congested Urban Road Networks. In *The 88<sup>th</sup> Annual Meeting of Transportation Research Board Compendium of Papers*. DVD-ROM. Transportation Research Board of the National Academies, Washington, D.C., 2009.
22. Waller, S., and A. Ziliaskopoulos. On the Online Shortest Path Problem with Limited Arc Cost Dependencies. *Networks*, Vol. 40, No. 4, 2002, pp. 216-227.
23. Zhang, X., and J. Rice. Short-Term Travel Time Prediction. *Transportation Research. Part C, Emerging Technologies*, Vol. 11, No. 3, 2003, pp. 187-210.