Spatio-temporal Analyses For Prediction Of Traffic Flow, Speed And Occupancy On I-4

Srinivasa Ravi Chandra Chilakamarri Venkata

University of Central Florida
SPATIO-TEMPORAL ANALYSES FOR PREDICTION OF TRAFFIC FLOW, SPEED AND OCCUPANCY ON I-4

by

SRINIVASA RAVI CHANDRA, CHILAKAMARRI VENKATA
B.E. (Hons) Birla Institute of Technology and Science
M.S.C.E (Transportation) University Of Central Florida

A dissertation submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Department of Civil, Environmental and Construction Engineering in the College of Engineering and Computer Science at the University of Central Florida Orlando, Florida

Spring Term
2009
Major Professor
Haitham Al-Deek
ABSTRACT

Traffic data prediction is a critical aspect of Advanced Traffic Management System (ATMS). The utility of the traffic data is in providing information on the evolution of traffic process that can be passed on to the various users (commuters, Regional Traffic Management Centers (RTMCs), Department of Transportation (DoT),…etc) for user-specific objectives. This information can be extracted from the data collected by various traffic sensors. Loop detectors collect traffic data in the form of flow, occupancy, and speed throughout the nation. Freeway traffic data from I-4 loop detectors has been collected and stored in a data warehouse called the Central Florida Data Warehouse (CFDW™) by the University of Central Florida for the periods between 1993 – 1994 and 2000 - 2003. This data is raw, in the form of time stamped 30-second aggregated data collected from about 69 stations over a 36 mile stretch on I-4 from Lake Mary in the east to Disney-World in the west. This data has to be processed to extract information that can be disseminated to various users.

Usually, most statistical procedures assume that each individual data point in the sample is independent of other data points. This is not true to traffic data as they are correlated across space and time. Therefore, the concept of time sequence and the layout of data collection devices in space, introduces autocorrelations in a single variable and cross correlations across multiple variables. Significant autocorrelations prove that past values of a variable can be used to predict future values of the same variable. Furthermore, significant cross-correlations between variables prove that past values of one variable can be used to predict future values of another variable. The traditional
techniques in traffic prediction use univariate time series models that account for
autocorrelations but not cross-correlations. These models have neglected the cross
 correlations between variables that are present in freeway traffic data, due to the way the
data are collected. There is a need for statistical techniques that incorporate the effect of
these multivariate cross-correlations to predict future values of traffic data.

The emphasis in this dissertation is on the multivariate prediction of traffic
variables. Unlike traditional statistical techniques that have relied on univariate models,
this dissertation explored the cross-correlation between multivariate traffic variables and
variables collected across adjoining spatial locations (such as loop detector stations). The
analysis in this dissertation proved that there were significant cross correlations among
different traffic variables collected across very close locations at different time scales.
The nature of cross-correlations showed that there was feedback among the variables, and
therefore past values can be used to predict future values.

Multivariate time series analysis is appropriate for modeling the effect of different
variables on each other. In the past, upstream data has been accounted for in time series
analysis. However, these did not account for feedback effects. Vector Auto Regressive
(VAR) models are more appropriate for such data. Although VAR models have been
applied to forecast economic time series models, they have not been used to model
freeway data.

Vector Auto Regressive models were estimated for speeds and volumes at a
sample of two locations, using 5-minute data. Different specifications were fit –
estimation of speeds from surrounding speeds; estimation of volumes from surrounding
volumes; estimation of speeds from volumes and occupancies from the same location;
estimation of speeds from volumes from surrounding locations (and vice versa). These specifications were compared to univariate models for the respective variables at three levels of data aggregation (5-minutes, 10 minutes, and 15 minutes) in this dissertation. For data aggregation levels of <15 minutes, the VAR models outperform the univariate models. At data aggregation level of 15 minutes, VAR models did not outperform univariate models. Since VAR models were used for all traffic variables reported by the loop detectors, this made the application of VAR a true multivariate procedure for dynamic prediction of the multivariate traffic variables – flow, speed and occupancy. Also, VAR models are generally deemed more complex than univariate models due to the estimation of multiple covariance matrices. However, a VAR model for $k$ variables must be compared to $k$ univariate models and VAR models compare well with AutoRegressive Integrated Moving Average (ARIMA) models. The added complexity helps model the effect of upstream and downstream variables on the future values of the response variable. This could be useful for ATMS situations, where the effect of traffic redistribution and redirection is not known beforehand with prediction models.

The VAR models were tested against more traditional models and their performances were compared against each other under different traffic conditions. These models significantly enhance the understanding of the freeway traffic processes and phenomena as well as identifying potential knowledge relating to traffic prediction. Further refinements in the models can result in better improvements for forecasts under multiple conditions.
ACKNOWLEDGMENTS

I would like to extend my heartfelt gratitude and appreciation for my advisor Dr. Haitham Al-Deek, for being extremely supportive during various phases of this challenging journey. Through all the challenges, his insight, and encouragement have given me the enthusiasm to aspire higher and be better. I would also like to extend my thanks to my Committee members – Dr. Essam Radwan, Dr. Amr Oloufa, Dr. Nizam Uddin and Dr. Shelby Gerking for graciously accepting to serve on my dissertation committee.

I would like to thank all the friends and colleagues and teachers and researchers, who always had something to teach me during my stay at UCF. I would like to thank Anurag, Aparna, Om, Swapna, Smitha, Emam, Hatem, Pradeep, Shankar, Fathy, Yan, Sai, Fakir, Raghu, Bharat, Yueliang, Piyush, Albindser, Nezam, Dixit, Xuesong, John, Jason, Rami, Taylor, Noor, Ali, Kirolos, Ana, Prem, Hani, Sid, Turen, Brenda, Olivia, who have made my stay as a student memorable. I would also like to thank Dr. Jack, Patrick, Dr. Balaji, Dr. Amr, who helped pique my interest in the research projects I worked at UCF. I would also like to thank my closest friends Ashis, Madhu, Mala, Srinivas, Raja Shekar, Nishanth, Kranthi, Veedee, Sandeep, Vidhya, Sailaja, Sasanka, and the whole “Kool-it group”; my friends from BITS – Abhishek, Santosh, Urvashi, Archana, Aparna, Slash and others; my school friends K.K.Chaitanya, Vikram, Ganesh, Pavani, Soumya, Caroline, Santhi Bhushan, and others, for having confidence in me and encouraging me to scale greater heights. Also I would like to express my heartfelt gratitude my surrogate family of Ravi Swapna, Vijay Sreevidya, Aparna Sridhar, and Kranthi Sandhya, whose support has helped me to get through the good and bad times. I
hope I have learnt all the right things from them, and will take something from their lessons to me wherever life takes me.

I would also like to thank all my teachers, right from my schooling days in India – Jayanthi Madam, Susheela Madam, Reva Madam, Vasantha Madam, Sebastian Sir, Vijay Anand Sir, Rachappa Sir, Ramesh Sir, Bhanu Chander Sir, Lalitha Madam, Vijay Lakshmi Madam, Prasurnamba Madam, Krishna Jayaram Madam, Sri Ram Sir, and all teachers from Mahatma Gandhi Memorial High School from Hyderabad; Krishna Moorthy Sir, Lakshmi Karra Madam, Anitha Madam, Subba Rao Sir, and all other teachers from Little Flower Junior College; Dr. Sarkar, Dr. S.P. Gupta, Dr. A.P. Singh, Dr. Rajiv Gupta, Mrs. Subramanian, Dr. Sangeeta Sharma and all teachers from BITS-Pilani, who have left indelible impressions on me.

I would like to thank my family the most – my parents Sri C.V. Sesha Narasimhacharyulu and Smt. C.V.M. Srivalli, my brother-in-law and my sister Sri. Girija Ramana Seshu and C.Padma Smitha, whose dreams and sacrifices are the reason behind how far I got today. I would like to thank all my uncles and aunts and cousins, and grandparents – whose wishes and blessings have helped me infinitely in this journey. I hope I make them all proud.

The greatest of all thanks should go to the Supreme Almighty God, who has made everything possible.
# TABLE OF CONTENTS

ABSTRACT ......................................................................................................................... ii
ACKNOWLEDGMENTS ................................................................................................. v
TABLE OF CONTENTS ................................................................................................. vii
CHAPTER 1. INTRODUCTION ......................................................................................... 1
  1.1 Motivation ............................................................................................................... 6
  1.2 Objectives .............................................................................................................. 7
  1.3 Organization of the Dissertation ........................................................................... 9
CHAPTER 2. LITERATURE REVIEW ............................................................................. 11
  2.1 Statistical Approaches .......................................................................................... 11
  2.2 State Space Models .............................................................................................. 14
  2.3 Non Parametric Models ....................................................................................... 16
  2.4 Soft Computing Approaches ............................................................................... 17
  2.5 Conclusions From Literature Review .................................................................. 19
CHAPTER 3. METHODOLOGY ...................................................................................... 20
  3.1 Univariate Time Series Analysis .......................................................................... 25
  3.2 Cross Correlations ............................................................................................... 29
  3.3 Multivariate Time Series Analysis ........................................................................ 31
CHAPTER 4. CROSS CORRELATION ANALYSIS SPATIAL TIME SERIES OF
FREEWAY TRAFFIC SPEEDS .................................................................................. 35
  4.1 Data Description .................................................................................................. 36
  4.2 Cross Correlation Analysis For Speeds And Volumes ............................................ 38
  4.3 Conclusions .......................................................................................................... 57
CHAPTER 5. MULTIVARIATE PREDICTION OF FREEWAY TRAFFIC SPEEDS
AND VOLUMES USING VECTOR AUTOREGRESSIVE MODELS ......................... 59
  5.1 Data Description .................................................................................................. 60
  5.2 Vector Auto Regressive Model ............................................................................ 61
  5.3 Results .................................................................................................................. 65
  5.4 Conclusions .......................................................................................................... 74
CHAPTER 6. MULTIVARIATE MODELS FOR FREEWAY TRAFFIC PREDICTION
AT THREE TIME SCALES ON INTERSTATE 4 ......................................................... 77
  6.1 Introduction .......................................................................................................... 77
  6.2 Data Description .................................................................................................. 78
  6.3 Methodology ........................................................................................................ 78
    6.3.1 Vector Auto Regressive Model .................................................................... 80
  6.4 Results .................................................................................................................. 82
  6.5 Conclusions .......................................................................................................... 105
CHAPTER 7. CONCLUSIONS AND FUTURE WORK ..................................................... 107
  7.1 Summary Of Research ......................................................................................... 108
  7.2 Recommendations For Future Work ................................................................. 111
APPENDIX A: SPEEDS SURROUNDING STATION 8 AND CORRESPONDING
CCF AND ACF ............................................................................................................. 113
APPENDIX B: VOLUMES SURROUNDING STATION 35 AND CORRESPONDING
ACF AND CCF ............................................................................................................. 115
CHAPTER 1. INTRODUCTION

Traffic processes arise out of trips made by travelers in their respective travel modes or vehicles for purposes of work, education, shopping, recreation and other miscellaneous activities. While activities related to work and education usually conform to standard time schedules, other activities are not generally time specific. The multitude of travel activities give rise to different traffic patterns that juxtapose at any moment in time resulting in the vehicular traffic that occupy land (streets and highways), water or air. The different modes making these trips, use access provided by the respective transportation infrastructures. At the same time, their presence on their respective infrastructures (land, water or air) can impede other vehicles' movements and access. Therefore, traffic processes over a given region can be thought of as the confluence of vehicles arising due to multitude of human activities, connecting different locations, providing a measure of access, as well as impedance.

With respect to land (road) networks, the limits on the availability of space for providing access between different locations and the added impedance of vehicles already accessing available space for their trips results in a non-zero, finite activity time for any additional vehicle making a trip. Land is constrained not only in the limits of its extent, but also in its use. Because almost all major human activity is concentrated on land, additional constraints are imposed on using land towards providing access. Roads have finite dimensions and finite capacities for vehicles. Therefore, trips on road networks take finite times. Added to that, different conditions on the road network, e.g. geometrics, access to the facility, vehicular capabilities, traffic conditions, driver
behaviors and characteristics, situational constraints, all have an effect on the travel times for the trips of the vehicles. Some of these conditions are dynamic and vary through different dimensions (time, space, demographics, etc.), some are random (incidents, etc.), which translate to an uncertain or partially deterministic and dynamic traffic, and by extension, travel time.

There has been an unmistakable upward trend in the usage of road networks over the last two decades. The statistics from the Federal Highway Administration (FHWA) reveal that between 1990 and 2007, about 4% more roads have been built, but there are about 50% more vehicles using them - and they add an additional 47% annual vehicle miles traveled (Highway Statistics 2007). The average annual delay per traveler increased from 27 hours in 1990 to 38 hours in 2005 (Urban Mobility Report, 2007). This spread of congestion is more significant in large metropolitan areas, but has been witnessed in the smaller areas too, albeit to a lesser extent. The congestion translates to a cost that is lost (due to loss of productivity), and has other environmental side-effects. To counter this loss, traffic planners figure that the best solution is to manage traffic flow more efficiently. Intelligent Transportation Systems (ITS) technologies offer intelligent solutions to managing traffic and can be efficient and economical alternatives to constructing more mileage of roads.

Freeway systems cater to a vast proportion of traffic demand, primarily because they are limited access, free and uninterrupted flow facilities. In Florida, the interstates run for about 1500 miles, carrying about 36,000 million annual vehicle miles per day (Highway Statistics 2007). Due to such huge demand, freeway systems are usually busy
and at peak hour demands in urban Central Business District (CBD) areas, the freeways suffer from severe congestion. Since freeways cover wide geographic and demographic extents, more often than not, the issues related to freeways assume county level or state level importance. Due to this, there is a continuous pressure to improve the freeway systems. Building roads is not always a feasible or an intelligent solution, and Intelligent Transportation Systems (ITS) and Advanced Traffic Management Systems (ATMS) technologies offer intelligent and optimal choices to improve the freeways. More often than not, these technologies provide information to the traveling public, which can influence their trip choices. The concentration of congestion has a spatial and temporal extent. If the right information is available through the ITS and ATMS technologies, travelers could be subtly persuaded to change their trip times or routes or modes that can “spread” the vehicles in the relatively unused parts of the road network. These technologies, in turn heavily rely on the existing traffic data and the underlying information that can be derived about the freeway systems. Typical traffic data includes speeds, counts, headways, occupancies, link travel times, etc.

A variety of intrusive and non-intrusive technologies can be used to gather traffic data. Intrusive technologies need installation on / within the pavement and have to disturb the traffic during their installation / maintenance to be able to gather data. Traditionally, loop detectors, which are intrusive technologies, have been employed very widely for traffic data collection. Video cameras, radio and microwave sensors are examples of non-intrusive detection technologies. While single loop detectors can provide volume (or vehicle count per time) and occupancy data (proportion of time the detector is occupied by vehicles), double loops – a pair of loops separated by a distance
typically about 20 feet, can provide speed measurements data as well. Loop detectors are traffic sensors that are embedded into the pavement and detect the presence of axles of vehicles. The data collected by loop detectors is useless by itself, unless it can contribute to further the understanding of the traffic phenomena. This understanding is important for developing an intelligent system that can provide real-time traffic and predict future traffic conditions and phenomena that can be utilized in an Advanced Traffic Information System (ATIS).

The current hardware and software capabilities permit the collection of real-time traffic data as well as its archival. When archived for sufficient time, this data becomes historical data that can provide some important insights into the evolution of the traffic processes. Therefore, traffic operators are faced with two possibilities for provision of information – provide real-time data continuously and / or provide predictive data based on the insights provided by the archived data. Currently, a significant proportion of traffic information providers display real-time information, through maps or by providing numerical information (segment-wise speeds / travel times / volumes) on Dynamic Message Signs (DMSs) installed at various locations on the expressway / freeway system. The Figure 1-1 given below shows two maps – the first one for a region in Central Florida, provided by Florida Department of Transportation, (through its transportation portal www.fl511.com), and the second one for Atlanta region, provided by Georgia Department of Transportation (through its transportation portal www.georgia-navigator.com). These maps show the real time conditions for these regions on the network.
Source: www.fl511.com

Source: www.georgia-navigator.com

Figure 1-1: Two sources with real-time traffic information maps (FDOT, G-DOT)
One of the limitations with provision of real-time data only is that the general public may not anticipate how situations would change or evolve in the short term. For example, a traveler might decide to make a trip at 5:30 pm, and notice that a segment \( x \) is coded blue (which could mean medium congestion), when he/she checks [www.fl511.com](http://www.fl511.com). However, when he/she reaches segment \( x \) 10 minutes later, he/she might find that the segment is a lot more congested than when he/she had accessed the site for information. Therefore, if the information system has the capability to anticipate the future, the system can provide more reliable information to the customer.

### 1.1 Motivation

With the deployment of loop detectors over the 36 mile length of I-4 (as of March 2003), the traffic data over that length has found its way to the Regional Traffic Management Centers (RTMC), and has been distributed to the various planning, administrative and research units. University of Central Florida had collected this data over the years. Currently, it houses the Central Florida Data Warehouse™ which holds the I-4 traffic data for almost a decade (with the exception of the years 1994 through 1997). This data has been used for separate but interrelated research on travel time algorithms, incident detection, data filtering, travel time reliability, crash analyses, crash predictions, etc. By its very nature, traffic data is correlated in time and space dimensions. As traffic enters the freeway at an on ramp, it moves with certain velocities on the freeway (depending on the conditions on the freeway) before it exits off an off-ramp. It is therefore, not unnatural to find some spatio-temporal correlations among the traffic data collected. By identifying the structure of these correlations among the data, it is expected that some discoveries about the relationships that were hidden in the data can be identified.
Ultimately, it is up to the users involved in the system (in this case, the managers of the transportation system and the public who use the system), who are empowered to make decisions based on choices available, to make improvements to the system.

In transportation systems and transportation management, the data that is most frequently involved are the freeway flows, traffic signal timings, travel times, accident data, etc. These data often reflect the state of the transportation system, and can serve as a guide to its management. But very frequently, it is not just a single attribute of data that can independently give valuable information. For example, if a freeway has very high flows, it does not reflect its performance in a significant way. However, if we know the relationship between flows and travel times (e.g., high flows and modest travel times), we have a better idea of the performance of the system. Therefore it is often essential to look at combinations of data and their attributes to get a comprehensive idea of the situation.

The specific emphasis in this study is on the prediction of the freeway traffic variables (flow, occupancy, speed) from loop detector data, using the under-utilized multivariate, to characterize spatial and temporal relationships.

1.2 Objectives

The primary focus of this proposal is to investigate the viability of spatial, temporal and the combination of spatial and temporal (spatio-temporal) techniques with respect to the I-4 traffic data warehouse at UCF. The spatial and temporal analysis techniques are expected to aid in the following objectives:
- Compute the correlations among the data to check the evidence of non-stationarities and feedback among the data.

- Fit univariate time series models to each of the de-trended and de-seasonalised variables.

- Perform multivariate analysis for one variable collected at different locations.

- Perform multivariate analysis for different variables collected at same location.

- Perform multivariate analysis for different variables collected at multiple locations.

- Examine the effect of temporal aggregation on the performance of different models.

- Predict the values of the traffic variables at locations for future times using the best model for each aggregation.

In the conventional set up, univariate Box Jenkins techniques for ARIMA model fitting, non-parametric models and neural networks have been used for the forecasting problem of the freeway traffic variables. The spatial component of the data has either been neglected or under-utilized as will be evident from the literature review. For a better prediction model, it is imperative that the variation of the traffic on spatial and temporal levels be taken into account that provides a more comprehensive and intuitive specification of the freeway traffic data generating mechanism. This dissertation aims to provide and prove the success of spatial and temporal techniques to achieve the aforementioned objectives, with particular emphasis on traffic prediction – prediction of traffic flow, speeds and occupancies. The success of any proposed procedure depends
significantly on the availability of data. The freeway traffic data is available, for a few years from the I-4 CFDW™. The models illustrated in this dissertation can be applied to different regions / freeways nationwide, as long as the data collected along these freeways conforms to a multivariate nature.

1.3 Organization of the Dissertation
The dissertation is organized into seven chapters. The description of these chapters is given below:

Chapter 1 provides the motivation, background and objectives of the research for this dissertation, the need for traffic prediction and the need to look into multivariate nature of traffic. Chapter 2 lists the literature review in the field of traffic prediction and the various approaches by multiple researchers to tackle prediction. Chapter 3 provides an elementary background in time series analysis, modeling, cross correlation analysis and multivariate time series estimation. Chapter 4, is derived from a paper published in Transportation Research Record journal by TRB, Washington DC., 2008, discusses the concept of cross correlation as it relates to traffic variables. It shows that univariate time series do not account for cross correlations in their formulation, and therefore cannot capture the system dynamics completely. Significant graphical evidence is provided to prove the same. Chapter 5 is derived from a paper accepted for publication in Journal of Intelligent Transportation Systems (J-ITS), 2008., and it provides the background on multivariate time series analysis and introduces Vector Auto Regressive (VAR) models for prediction of speeds and volumes from the neighboring conditions. The comparison between univariate and multivariate prediction models is provided. Chapter 6, is derived from a paper submitted to ASCE journal, and is currently under re-review, and it
examines the performance of VAR models vis-à-vis univariate models for prediction of volumes at three different data aggregation levels – 5 minutes, 10 minutes and 15 minutes. It comments on the applicability and success of VAR models compared to univariate models, and provides pointers to when these competing models can be applied successfully. It shows that the performance of VAR models marginally declines at higher time scales. Chapter 7 summarizes the contribution of this dissertation and lists the conclusions and directions for future research.
CHAPTER 2. LITERATURE REVIEW

Traffic data from the freeways has been subjected to extensive analysis for traffic prediction, incident detection, imputation of traffic data, etc. The techniques involved in traffic data analysis ranged from simple averaging to sophisticated statistical techniques to several artificial intelligence techniques. As far as prediction is concerned, most of these analyses modeled correlation of data along time domain, while correlation along the spatial dimension has been traditionally ignored or dealt on a very limited basis. The following sections provide a literature review in forecasting freeway traffic.

2.1 Statistical Approaches

With the advent of extensive traffic data gathering mechanisms and the emphasis on data archival, the Traffic Management Centers (TMCs) across the nation and around the world have deployed loop detectors. Typically, these loop detectors are embedded into the pavement, one in every half mile section. They have the capability of inferring vehicle presence, vehicle count and under a special configuration (two closely spaced loop detectors known as dual loops or speed traps) can estimate the traffic speed, by the process of magnetic induction. Once they are installed, they do not obstruct the traffic in any way and can report the data to the TMCs at set intervals – typically set at 20 s or 30 s intervals. This makes the data from each loop detector a realization of time series that can be modeled by the extensive univariate time series methods as prescribed by Box and Jenkins (1970).

Ahmed and Cook (1979) used the time series methodology as described by Box and Jenkins techniques to freeway traffic data. They fitted an Auto Regressive Integrated Moving Average (ARIMA) model to the traffic data and compared it with simple moving
average and exponential smoothing models. They found that ARIMA model defined as ARIMA (0, 1, 3) model was a better prediction model than the others. The parameters of this model describe the number of lags (or past observations) that contain information about a current observation. This notation will be explained in detail in the next chapter. Nihan., et al., (1980) illustrated a similar use for Box and Jenkins technique in traffic forecasting.

Danech-Pajouh and Aron (1991) developed “ATHENA”, a program that used a layered statistical approach to group the data into well defined clusters and apply linear regression on them. They showed that the ATHENA model was superior to the other methods, though the complex parameterization was a downside for transferability of the model.

Hobeika and Kim(1994) used multiple regression models to forecast traffic from historical averages at a location and the current conditions on the upstream links. They used 15 minute traffic flows and showed that their technique provided better forecasts than a historical average model.

Williams, (1998) had shown that the Seasonal Autoregressive Integrated Moving Average (SARIMA) time series model is a suitable model for forecasting 15 minute traffic flow. The SARIMA time series model used the Box-Jenkins method of identification, estimation and forecasting, and used a parsimonious 3-parameter model for predictions. Williams and Hoel (2003) used SARIMA to forecast 15 minute traffic flows on two freeways and compared it with naïve benchmark forecasts from random walk forecasts and historical averages. They concluded that the parsimonious 3 parameter SARIMA model should be a benchmark for comparison with other models.
Shekhar (2004) proposed the use of adaptive filtering techniques (state space methods) for adaptive and recursive estimation of the SARIMA model parameters as proposed by Williams (1998). While the traditional model needs a block and offline estimation technique to estimate its parameters, the adaptive algorithms can adapt the model to the changing conditions online.

Non-linear time series analysis techniques for traffic predictions were explored by D’Angelo et al., (1999), Nair et al., (2001), among others. They assumed that traffic flow was deterministic but exhibited random-like behavior due to chaos in the system.

Yue (2006) also observed cross correlations for volumes from a street network in Hong-Kong between a location and it’s upstream. However, they neglected downstream influence. Williams (2001) evaluated the use of ARIMAX model using the upstream flow data to a location as the input series to predict the flow at a location 30 minutes in to the future. Since the upstream location was about 90 kms distant from the location under study, the effect of downstream location was neglected. Kamarianakis and Prastacos (2002) used space-time modeling of traffic flow on urban arterial networks in Athens, Greece, by defining neighborhood matrices that reflect the position of the loop detectors and incorporating it in their model specification. Their results reflected the need for different models for different times of the day, as well as the need to consider the multivariate nature of traffic over spatial domain.

These statistical approaches have concentrated on traffic predictions in the network by using univariate time-series models. They either do not acknowledge the spatial component or under-utilize the spatial component of the traffic, by considering only the upstream locations.
2.2 State Space Models

Okatuni and Stephanedes (1984) first applied Kalman filter theory to the prediction of traffic flows. The variable to be predicted (traffic flow) was unobserved variable, and the covariates to predict the traffic flow that were measured (or observed) were defined as the state variables. These quantities were related equations that related the unobserved variable to the current state of the observed variables and a state equation that defined how the state changes through time (usually modeled as a first order autoregressive process). By the method of recursive estimation, the parameters of the state space model were estimated and were used to generate forecasts for $k$ steps ahead.

Jiang (2002) used Kalman predictor using Kalman filtering, by filtering with first order ARMA process to dynamically predict the traffic flows and onset of congestion in construction zones. The author found that Kalman predictor combined with the AR(1) model was more dynamic and accurate when compared to the pure time series model.

Stathopoulos and Karlaftis (2001) explored the application of frequency domain techniques in assessing the effect of neighboring location traffic data and their dependencies on each other. By using spectral and cross spectral analysis on 3-minute traffic flows on two specific adjacent loop detector stations, the researchers proved that there was a positive autocorrelation between time lags of 1-2 time slices. They also proved that the short and long term traffic fluctuations were highly correlated between the two stations – thus proving that most of the traffic flows over both the detectors with some losses – due to deceleration. The authors therefore recommended the development of parsimonious multivariate state space models for prediction.
Yu (2004) proposed another hybrid model with time series using a hidden variable to denote the state of the system to denote different traffic conditions. By specifying different transition probabilities for different state transitions and defining different ARIMA models for different states, the forecast was defined as a function of the (future) state of the variable. The future state was estimated by the transition probabilities.

Stathopoulos and Karlaftis (2003) then used multivariate approaches using upstream detectors to forecast the traffic flows in the downstream section for an arterial network in Athens, Greece. They applied kalman filters for state space equations and extended it to use the cross correlations of from the upstream detectors. Using 3-minute data urban arterials, the authors developed models that used data from upstream detectors to improve on the predictions of downstream location. Their results reflected the need for different models for different times of the day, as well as the need to consider the multivariate nature of traffic over spatial domain.

Yang et al., (2004) proposed an on-line adaptive state space model by taking historical off-line data into account. A recursive algorithm was used to obtain the computational efficiency as well as reduced storage. A maximum likelihood estimate of the noise covariance matrix and transition coefficients matrix was provided off-line and optimal time variant parameters were calculated on-line. The authors showed that the state space model with the nonzero noise covariance matrix outperformed other algorithms with the loop detector data on I-405, California.

Most of the studies described earlier relied on univariate statistical and time series models. These studies modeled only one variable – flow, across a single dimension – time. However, it has been acknowledged in various works that the spatial component induces some variability in the traffic patterns that evolves through time. Also only a
single upstream loop detector station has been considered to improve the prediction for the downstream station.

### 2.3 Non Parametric Models

Davis and Nihan (1991) used non-parametric regression to forecast short-term freeway traffic flows, by using the $k$-Nearest Neighbor formulation ($k$-NN). Freeway loop data was used to test the $k$-NN approach and to compare it with time-series predictions. The $k$-NN method did not outperform time-series forecasts, but was comparable.

Smith, et al., (2001) successfully applied a variant of the non-parametric regression known as the approximate nearest neighbor non parametric regression to arterial traffic data in Virginia and could predict accurately. Their method was similar to Case-Based Reasoning (CBR), which searches historical data for patterns that are “closest” (or nearest neighbors) to the current data and based their forecasts for the current data on the future “forecasts” of the neighbors.

Clark (2003) proposed the application of multivariate non-parametric regression models. This reflects the multivariate relationships within traffic variables, and provides an intuitive expression for the forecasting models of traffic flow.

Sun et al., (2003) applied “local linear regression model” to forecast traffic data. In this methodology, the number of covariates is defined by the “bandwidth” matrix, and it forms a part of the objective function (loss function) that has to be minimized and is reported to be superior to nearest neighbor approaches.

Turochy and Pierce (2004) used Multivariate Statistical Quality Control (MSQC) measures with $k$-nearest neighbor non-parametric regression to determine the “normality” or “abnormality” of the current observation when compared to historical conditions.
(historical means and variances) for freeway traffic loop detector data. This “normality” of the observation was used to improve the forecast.

2.4 Soft Computing Approaches

Taylor (1994) used Artificial Neural Networks (ANN) for freeway traffic data prediction and ramp metering. By training the ANN with recent samples of traffic volume and occupancy at a station, combined with its upstream station data, weekday traffic flow was successfully predicted.

Van Der Voort et al., (1996) used a class of neural networks known as the self organizing neural networks developed by Kohonen (1995) – referred in literature as Self Organizing Maps (SOMs) or Kohonen networks. The SOM was used to cluster traffic data and an ARIMA model was developed for each cluster. This technique was comparable to ATHENA models in its accuracy.

Kirby, et al., Van Arem, et al., (1997) compared different Neural Networks formulations with ARIMA models for forecasting. For forecasts ranging from 30 minutes to 2 hours ahead, ARIMA and ATHENA models outperformed neural networks.

Abdulhai et al., (1999) used neural network architecture with time delay – and concluded that spatial contribution was critical for prediction up to 15 minute flows, and 3 stations upstream and 3 stations downstream are essential for forecasting at those temporal levels.

Chen and Miller (2001) used Resource Allocating Network (RAN) architecture of neural networks - it calculated the posterior probability of a function given the prior probability and a new observation by recursive non-linear least squares algorithm. This network was then trained and used for forecasting. The simple dynamic neural network
with five hidden units performed the best, and could learn the incident patterns also by the application of piece-wise models.

Yin et al., (2002) proposed a fuzzy neural architecture for urban traffic flow prediction. The architecture consists of two modules – one that clustered the input patterns; and other, that based its predictions on the patterns.

Lin (2001) proposed a new model based on the Gaussian Maximum Likelihood (GML) method. The two key variables in the GML model were traffic flow and traffic flow increments in time intervals of 5 minutes. Normal distribution for the two variables was assumed. An estimate of the predicted flow was derived by maximizing the likelihood of the flow level at the next time interval and was expressed as a product of the two probability density functions.

Tang et al., (2003) compared short term predictions of AADT with four approaches - an ARIMA(1,0,0) model; a Neural Network model with two input units, a nearest neighbor non-parametric model, and a GML model as proposed by Lin (2001). The authors proved the superiority of the non-parametric model as well as the GML model over the time series and neural network model for AADT forecasting. However, the authors conceded that the validity of assumptions for the two key random variables in the GML formulation might be suspect.

Xie et al.,(2006) proposed the application of the Kalman filter with discrete wavelet analysis in short-term traffic volume forecasting. Discrete wavelet decomposition and reconstruction analysis was used to divide the original data into several approximate data such that the Kalman filter could be applied to the de-noised data only and the prediction accuracy can be improved. Traffic volume data collected from Interstate 80 was used in this study. The test results indicated that the wavelet
Kalman filter model consistently performed better than the Kalman filter model in terms of accuracy and stability.

2.5 Conclusions from Literature Review

From the literature review, it is evident that some of the most promising techniques for traffic prediction can be broadly divided into ARIMA models, State Space models, Non-parametric models and Neural Network models. In all these techniques, the effect of multivariate correlations in traffic data has been ignored. While non-statistical approaches have been “inclusive” of attribute neighborhood, and spatial neighborhood, statistical techniques have been strongly led by ARIMA formulations. While being intuitive as well as accurate, univariate models can be improved by considering the multivariate correlations in traffic data. Non-parametric approaches and Neural Networks do not have the power of interpretability that is provided by the statistical approaches. The interpretability of parameters provides a useful tool for engineers and operators to envision alternate scenarios and forecast the system wide impacts of these scenarios (however big the system may be). However, a framework needs to be available to assess the effects of different traffic variables on each other that are in close proximity – both in space and time. The rest of this dissertation proves that there is predictive information available from variables that are related to each other in time or space.
CHAPTER 3. METHODOLOGY

In the context of traffic prediction and traffic information, different stakeholders have different information demands. For traveling public, the pertinent information is travel time, which is the most important measure. Other information is usually derived from travel time (from the traveler’s perspective) and this includes delay and congestion, mean travel times, travel time reliability, and so on. For traffic planners, the critical information is demand and available capacity. For traffic engineers, efficient operations being the goal, the pertinent information relates to system wide or network wide measures of delays, Level of Service (LOS) for corridors and networks, and demand management. With the advent of ITS and ATMS strategies, there is an increasing demand for efficient and extensive data collection strategies that can provide relevant information to all stakeholders involved (e.g., Advanced Traveler Information Systems – ATIS).

Within this context, the information needs require data to be collected. As was explained in earlier sections, there are certain traffic variables that can be directly measured, while certain variables can be derived. For example, demand can be measured by the number of vehicles traveling in a certain amount of time. Travel time can either be measured directly, using specialized vehicles (probe vehicles), or determined indirectly by measuring average speeds of vehicles over a segment or a point. The mean and variations of these quantities provide further information that can be assessed by different stakeholders.

It is therefore possible to provide useful information by collecting some rudimentary data and then utilizing calibrated mathematical models to derive other
quantities. Loop detectors are prevalent on most freeway systems in the US, providing data on the following quantities:

- **Volumes:** Number of vehicles traversing a point in a certain time interval
- **Occupancy:** The proportion of time a vehicle occupies a small section of the roadway.
- **Speed:** The average speed of vehicles passing over the detectors.

Since vehicles traveling at slower speeds tend to spend greater time traversing small sections of road, occupancy was also used as a proxy measure of density or congestion. The dual loop detectors which are installed on I-4 can measure speed directly.

The detectors collect data at spots that they are installed at discrete intervals of time. Typically these intervals vary between 20 seconds to 60 seconds. The locations and times of data collection are indices of traffic data from loop detectors. All relationships deciphered from traffic data will make use of these indices. In the prediction domain, these spatial and temporal assume importance, as information for future times needs to be forecast at different locations. The locations of a subset of dual loop detectors installed in the Central Florida region are provided in Figure 3-1.
The loop detectors are spaced at approximately half mile distance from each other. In the CFDW™ data-warehouse, data was available for 69 stations, in both east bound and west bound directions, starting from station 2 near Disney in the west, through station 71 near Lake Mary in the east (there is no location designated as station 39). At each station, there is a detector in each lane in each direction. In the west end of the corridor, stations 2, 3, 4, 5 have detectors reporting data for two lanes in each direction, while the rest of the stations reported for three lanes in each direction. The data collected by each detector was the speed, volume and occupancy for every 30 second interval. These data are very noisy and have a lot of variability and they can be averaged at various discrete intervals for different applications. Typically, 5 minute and 15 minute
aggregated data are used for real-time traffic information purposes. The speeds, volumes and occupancies at 30 second intervals for a typical day are shown in Figure 3-2. These quantities are aggregated as given by the following formulae:

\[
V_f = \sum_{i=1}^{n} v_{30}^i, S_f = \frac{\sum_{i=1}^{n} (s_{30}^i \times v_{30}^i)}{V_f}, O_f = \frac{\sum_{i=1}^{n} o_{30}^i}{n}
\]

Speeds are weighted by volumes, as simple averages will not capture the variation caused due to different volumes in a 5 minute interval. Where \( V_f, S_f \) and \( O_f \) are volume, speed and occupancy respectively, at aggregation level \( f \) (\( f=1 \) minute, 5 minutes, 15 minutes, etc), \( v_{30}^i, s_{30}^i, o_{30}^i \) are 30 second volume, speed and occupancy respectively, \( i \) is an index, and \( n \) is the number of discrete 30 second intervals at an aggregation level \( f \). For \( f=1 \) minute, \( n = 2 \); for \( f=5 \) minutes, \( n=10 \) and so on. The data is archived at 30 seconds and 5 minutes, and data quality is checked by imputation algorithms given by Al-Deek (2003).
30 second speeds for a typical day at loop detector on left lane

30 second volumes for a typical day at loop detector on left lane

30 second occupancies for a typical day at loop detector on left lane

Figure 3-2: 30 sec speeds, volumes and occupancies on left lane for station 31
### 3.1 Univariate Time Series Analysis

Let us consider a variable \( Y_t = \{y_1, y_2, y_3, \ldots, y_t, \ldots, y_T\} \), or denoted concisely by \( Y_t = \{y_t\} \) or simply \( y_t \). This denotes a variable that is measured along the time index \( t = 1, 2, \ldots, T \), where \( T \) is the length of the series. The time indices are assumed to be discrete and equally spaced intervals. Let \( y_t \) assume random values at each of the time indices \( t \). \( y_t \) would then be a random variable, following some probability distribution. In general, the statistical properties of random variables are determined by their moments. The first moment is the mean or expectation of the variable denoted by \( E(y_t) \) (which is equal to \( \mu \)), and the second moment is the covariance, denoted by \( E((y_t-\mu)(y_s-\mu)) \), where \( t, s \) are time indices \(< T \). Higher order moments also exist, and the complete specification of moments determines the probability distribution of the variable. If the statistical properties (or moments) for the random variable \( y_t \) are independent of \( t \), then the index does not provide any information about the variable itself. The ordering of \( y_t \) according to increasing \( t \) will not change any statistical property of \( y_t \). Such a random variable is called independent and identically distributed (i.i.d). When the variable follows a normal distribution, the iid property ensures that the inferences and hypothesis testing for random variable and the confidence intervals are valid.

However, when variables are sampled at close, regular time intervals, it is possible that closely sampled values contain some information about the “next” value of the variable. In a sense, the time index \( t \), can provide some information about what value can be expected of \( y_t \). In this case, each value of \( y_t \) at any time \( t \) is not random, but is dependent on \( t \), and implicitly, by its position in the series. As an example, each of the series in Error! Reference source not found., seems to show a persistent kind of behavior where on an average, large values tend to be followed by large values and smaller
values tend to be followed by smaller values. This behavior might not continue for long periods, otherwise the variable would indefinitely increase out of bounds. However, such behavior is the basis of time series analysis, which exploits the relationship between the random variable at different time indices, particularly towards forecasting future values ($t > T$).

A time series is “stationary” if the statistical properties (moments) are same for all values of $t$. In essence, a time series can be considered weakly stationary if the first and second moments of the series are not dependent on $t$. It means that it doesn’t matter when the observation is made, as we can expect the same statistical properties at all times. The first moment is the mean function $E(y_t) = \mu_t$. If $\mu_t$ is constant for all $t$, the subscript can be dropped and can be denoted as $\mu$. The second moment is $E((y_t-\mu)(y_s-\mu))$, where $t$, $s$ are time indices < $T$, denoted by $cov(y_t,y_s)$. The stationarity condition requires that $cov(y_t,y_s)$ depends only on the separation of $t$ and $s$ in time domain or $|t-s|$ denoted by $h$, known as lag, and not on $t$ or $s$. Therefore $cov(y_t,y_s) = \gamma_y(h)$, a function of $h$ only. This function is called the auto-covariance function. When $h=0$, this is variance, denoted by $\gamma_y(0)$, and the ratio of covariance at lag $h$ to the variance of the series is called autocorrelation (or self correlation), denoted by $\rho(h)$. By definition, $\rho(0)=1$. The sample versions of auto covariance and auto correlation are respectively:

$$\hat{\gamma}_y(h) = \frac{1}{T} \sum_{t=1}^{T-h} ((y_t-\bar{y})(y_{t+h}-\bar{y}))$$

$$\hat{\rho}(y_t, y_{t-h}) = \frac{\sum_{t=h}^{T} (y_t-\bar{y})(y_{t-h}-\bar{y}_{t-h})}{\sqrt{\sum_{t=1}^{T} (y_t-\bar{y})^2 \sum_{t=2}^{T} (y_{t-h}-\bar{y})^2}}$$
where $\hat{\gamma}_y(h)$ is the estimated sample auto covariance, $\bar{y}$ is the mean of time series, $h$ is the lag where the auto-covariance is to be calculated, $T$ is the length of the time series.

A stationary process is a series which does not have any long term trend or seasonality (which are functions of time index $t$), is characterized by fast decaying autocorrelation function. When the autocorrelation doesn’t decay quickly enough, certain transformations need to be applied to the data to ensure stationarity. Usually, visual inspection provides a good idea about stationarity. If the non-stationarity is due to a trend in the data, simple differencing can eliminate the trend. A simple difference is denoted by $\nabla$ operator and is defined as the difference between the current value of the variable with its immediate predecessor or lagged value $\nabla y_t = y_t - y_{t-1}$. This can also be represented as $\nabla y_t = y_t - B(y_t) = (1-B)y_t$ and $\nabla^d y_t = (y_t - B(y_t))^d$ where $B$ is the backshift operator, which lags the variable by one time step and $d$ is the order of differencing. The backshift operator can shift a variable $s$ time steps, so that $B^s(y_t) = y_{t-s}$ and $\nabla_s = (1-B_s)y_t = y_t - y_{t-s}$. A single difference operator can remove trends and a difference operator at lag $s$ can remove seasonality (of period $s$). A successive single and periodic difference operator can remove trends and seasonalities.

Once the non-stationary data is de-trended and de-seasonalized, the autocorrelations decay quickly. In this case, for the identification of time series models, an examination of both auto correlation and the partial autocorrelation function is required. Partial autocorrelations measure the effect of the variables at a specific lag, after the effects of intermediate lags are accounted for. Both of these correlation functions are examined to identify the univariate time series model. If the autocorrelation function
decays exponentially and the partial autocorrelation function disappears after a certain lag \( p \), then an Auto-Regressive model of order \( p \) (AR(\( p \)) model) is appropriate for the stationary data. If the autocorrelations disappear after a lag \( q \) and the partial autocorrelations decay exponentially, then a Moving Average model of order \( q \) (MA(\( q \)) model) is appropriate. If both autocorrelations and partial autocorrelations decay exponentially, an Auto Regressive Moving Average (ARMA(\( p,q \)) model) might be appropriate. Once the orders of model are selected, least squares or maximum likelihood methods can be used to estimate the parameters of the model. A general ARMA (\( p,q \)) model is specified as

\[
x_t = \nu + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}
\]

A pure AR model would have \( \theta_i = 0 \) (\( i = 1 \ldots q \)) and a pure MA model would have \( \phi_i = 0 \). \( x_t \) is the stationary series derived from \( y_t \). If \( y_t \) is stationary, then \( x_t = y_t \). It needs to be understood that there is no unique model identification strategy. Very often, there are different candidate models that appear appropriate to the data in question. A selection can be based on different information criteria that award better performance and penalize the model complexity in terms of parameters to be estimated. Most of these information criteria are based on the log-likelihood and a penalty function for number of parameters to be estimated. Two of the widely used information criteria are Akaike Information Criterion (AIC), Schwarz Bayesian Criterion (SBC).

\[
AIC = 2k - 2ln(L)
\]

\[
SBC = k \cdot ln(n) - 2ln(L)
\]

Where \( ln(L) \): log-likelihood function

\( k \): number of parameters to be estimated

\( n \): sample size
Once the model is identified, diagnostic checks need to be performed on the residuals to make sure that there are no auto correlations among the residuals, and that they are normally distributed. Under some stability conditions, the ARIMA model has infinite order specification of either pure AR terms (causal model) and / or pure MA terms (invertible). Further details can be inferred from Brockwell and Davis (2002).

3.2 Cross Correlations

The formula for calculating the cross correlation function for two series $Z_t$ and $X_t$ at a given lag $h$ is

$$
\rho(z_t, x_{t-h}) = \rho_{zx}(h) = \frac{\sum_{i=1}^{n-h} (z_i - \bar{z})(x_{i-h} - \bar{x})}{\sqrt{\sum_i (z_i - \bar{z})^2} \sqrt{\sum_i (x_{i-h} - \bar{x})^2}} = \frac{\text{cov}(Z_t, X_{t-h})}{s_z s_x}
$$

Similar to autocorrelation, both the series $Z_t$ and $X_t$ are assumed to be stationary. If there is non-stationarity which is evidenced by a linear or non-linear trend in the series, or a varying variance, the series need to be transformed to stationarity by differencing (to remove trends) or by a transformation (box-cox, logarithmic, etc). It has to be noted that in general $\rho_{zx} \neq \rho_{xz}$.

The cross correlation measures the strength and direction of the relationship between the input and response series. $\rho_{zx}$ is the effect of past values of $x$ on the current value of $z$. $\rho_{xz}$ is the effect of past values of $z$ on the current values of $x$. If the response series also influences the input series, then there is feedback in the system. In the cross correlation plot, correlations on the positive lag side indicate the effect of input series $x$ on $z$.

For traffic data, it has been established that the traffic flows downstream from upstream, with some time lag. Hence, it is reasonable to assume that the cross correlation
plots will show significant spikes for positive lags for the cross correlations of traffic for a location with its upstream station. If there is a significant spike at the negative lag, it means that the location in question affects the upstream station also.

In the literature there has been little work on explaining the cross correlations for traffic data between different sites. At most the cross correlation between a location and the traffic upstream has been explored. However, in freeway traffic due to congestion, the downstream traffic conditions also exert an influence on the traffic conditions on a location. If this is true, then an examination of the cross correlation function should reflect this fact. Let $s(x,t)$ be the time series of a traffic variable $s$ at location $x$ and time $t$. The location $x^+$ is a downstream location and $x^-$ is an upstream location to $x$. The lag $h$ is positive for cross correlation between $s(x,t)$ and $s(y,t-h)$ and negative for cross correlation between $s(x,t)$ and $s(y,t+h)$ where $y$ denotes either upstream or downstream location. The cross correlation coefficient is positive when the spike is above the horizontal axis and negative when below the horizontal axis. The band around the horizontal axis is the 95% confidence band. Any spikes beyond this band are significant. Then, we have

- If $\rho_{s(x,t)s(x^-,t-h)} = \rho^+(h)$ is significantly different from 0 for some $h > 0$, then past values of traffic at upstream location have some influence on the current values on traffic at the location $x$.
- If $\rho_{s(x,t)s(x^-,t+h)} = \rho^-(h)$ is significantly different from 0 for some $h > 0$, then current values of traffic at location $x$ have some influence on the past values on the upstream location.
- If $\rho_{s(x,t),s(x',t-h)} = \rho^*(h)$ is significantly different from 0 for some $h > 0$, then past values of traffic at downstream location have some influence on the current values on traffic at the location $x$.

- If $\rho_{s(x,t),s(x',t+h)} = \rho^*(-h)$ is significantly different from 0 for some $h > 0$, then current values of traffic at location $x$ have some influence on the future values on the downstream location.

A univariate time-series model for a traffic variable can be modified by the inclusion of past values of upstream / downstream variables if the past values of the upstream / downstream variables have an influence on the current location.

The interpretation of cross correlation function between two variables requires one of the variables to be designated as response and the other one as input. Moreover, the examination of cross correlation requires both the positive and negative lags has to be examined. A significant correlation on the positive lags is interpreted as the effect of past values of the input (the number of past values indicated by the highest lag of the significant correlation coefficient) on the future values of the output. However, if there are significant correlations on the negative side, it means that the past values of the response influence the future values of the input. This is called feedback. Transfer function models model the effect of past values of input on the response, but cannot handle feedback, and multivariate analyses in the form of Vector Auto Regressive (VAR) models are necessary. Brockwell and Davis(2002) can be referred for further details.

### 3.3 Multivariate Time Series Analysis

Let $Y = (y_1, y_2, ..., y_i, ..., y_T)$, where each $y_i$ is a vector of $K$ variables, and $T$ is the length of the time series. $Y$ is then a $(K \times T)$ matrix of $K$ correlated variables. Each variable can be
analyzed separately using univariate techniques discussed in the previous section or the whole $Y$ matrix can be analyzed as a time series variable (with stationarity conditions as explained earlier), but the analysis will be multivariate, taking inter-variable correlation into account. Therefore, the covariance matrix for $Y$ will comprise of covariance submatrices each variable with every other variable. This gives rise to cross-covariances between variables in addition to auto covariances for each variable. Let $\{i, r \in 1, \ldots, K$ and $j, s \in 1, \ldots T\}$, where $i, r, j, s$ index the observations in $Y$. The possible covariances among different observations in $Y$ are:

$E(y_{ij}, y_{is}) = \text{autocovariance of } i^{th} \text{ variable for lag } = |s-j|$. If $s = j$, it is the variance of $i^{th}$ variable.

$E(y_{ij}, y_{rj}) = \text{covariance between variable } i \text{ and variable } r \text{ at time } j$

$E(y_{ij}, y_{kl}) = \text{cross-covariance between variable } i \text{ and variable } k \text{ for lag } = j-l$

Suppose that $Y$ is a multivariate Vector Auto Regressive model, where each variable can be expressed as a linear combination of past values of other variables. Then, we can write

$$y_t = \nu + A_1 y_{t-1} + \ldots + A_p y_{t-p} + u_t$$

Where $A_1, \ldots, A_p$ are $K \times K$ coefficient matrices $u_t$ is the vector of uncorrelated white noise of dimension $K$, and $\nu$ is the vector of intercepts (of dimension $K$). These coefficient matrices can be estimated using Ordinary Least Squares or Maximum Likelihood Estimation.

Let us define

$$B = (\nu, A_1, \ldots, A_p) \text{ (dimension: } K \times (Kp+1))$$
\[
Z_t = \begin{bmatrix}
1 \\
y_t \\
\vdots \\
y_{t-p+1}
\end{bmatrix} \quad \text{(dimension: ((Kp+1)×1))}
\]

\[
Z = (Z_0, \ldots, Z_{T-1}) \quad \text{(dimension: ((Kp+1)×T))}
\]

\[
U = (u_t, u_{t+1}, \ldots, u_T) \quad \text{(dimension: K×T)}
\]

\[
y = \text{vec}(Y) \quad \text{(dimension: KT×1)}
\]

\[
\beta = \text{vec}(B) \quad \text{(dimension: ((K^2p+K)×1))}
\]

\[
u = \text{vec}(U) \quad \text{(dimension: (KT×1))}
\]

where \(\text{vec}()\) function is the column stacking function for a matrix. If \(X = \)

\[
\begin{bmatrix}
x_{11} & \cdots & x_{1C} \\
\vdots & \ddots & \vdots \\
x_{R1} & \cdots & x_{RC}
\end{bmatrix}_{R×C}
\]

then \(\text{vec}(X) = \)

\[
\begin{bmatrix}
x_{11} \\
\vdots \\
x_{R1} \\
x_{12} \\
\vdots \\
x_{R2} \\
\vdots \\
x_{RC}
\end{bmatrix}_{RC×1}
\]

Then the vector autoregression can be represented as

\[
Y = BZ + U
\]

Or by applying the vec() operator on both sides of the equation \(\text{vec}(Y) = \text{vec}(BZ) + \text{vec}(U)\)

\[
= (Z \otimes I_K)\text{vec}(B) + \text{vec}(U)
\]

which can be written as

\[
y = (Z \otimes I_K)\beta + u
\]

where \(\otimes\) is the Kronecker product. If \(C\) and \(D\) are two matrices defined as
Then the covariance matrix of \( \mathbf{u} \) denoted by \( \Sigma_u = \Sigma(\text{vec}(\mathbf{u})) = \mathbf{I}_q \otimes \Sigma_u \)

The multivariate version of Least Squares estimation would find \( \mathbf{\beta} \) to minimize

\[
S(\mathbf{\beta}) = \mathbf{u}'(\mathbf{I}_q \otimes \Sigma_u)^{-1}\mathbf{u} \\
= \mathbf{u}'(\mathbf{I}_q \otimes \Sigma_u^{-1})\mathbf{u} \\
= (\mathbf{y} - (\mathbf{Z} \otimes \mathbf{I}_K)\mathbf{\beta})'(\mathbf{I}_q \otimes \Sigma_u^{-1})(\mathbf{y} - (\mathbf{Z} \otimes \mathbf{I}_K)\mathbf{\beta}) \\
= \text{vec}(\mathbf{Y} - \mathbf{BZ})'(\mathbf{I}_q \otimes \Sigma_u^{-1})\text{vec}(\mathbf{Y} - \mathbf{BZ}) \\
= \text{tr}[(\mathbf{Y} - \mathbf{BZ})'(\mathbf{I}_q \otimes \Sigma_u)^{-1}(\mathbf{Y} - \mathbf{BZ})]
\]

By equating \( S(\mathbf{\beta}) \) to zero to find the minimum, we have

\[
\hat{\mathbf{\beta}} = ((\mathbf{Z}\mathbf{Z}')^{-1} \otimes \Sigma_u)(\mathbf{Z} \otimes \Sigma_u^{-1})\mathbf{y} \quad (\text{Luktepohl, 1993}) \\
= ((\mathbf{Z}\mathbf{Z}')^{-1} \mathbf{Z} \otimes \mathbf{I}_K)\mathbf{y}
\]

Further details are available from Luktepohl (1993).
CHAPTER 4. CROSS CORRELATION ANALYSIS  
SPATIAL TIME SERIES OF FREEWAY TRAFFIC SPEEDS

Short term traffic forecasting is the prediction of traffic variables up to future periods – typically less than an hour. This data collected by different traffic sensors (e.g., loop detectors) provides insight into the repetitive patterns in traffic that can be utilized towards prediction / forecasting of traffic data. Historically, there have been a) statistical approaches, b) soft computing approaches, c) non-parametric approaches to predict traffic on arterials and freeways. However, univariate statistical time series models have neglected the spatial relationship among different variables collected from different sites. This can be seen as a limitation of the traditional univariate time series models as the traffic data from neighboring sites can provide information that can be useful for prediction of traffic at any given site. Soft computing approaches like Artificial Neural Networks, Genetic Algorithms, Fuzzy systems, etc., do not have the power of interpretability and as it is difficult to isolate and interpret the effects of the dependent variables.

This chapter deals with showing the effect of the traffic conditions around the neighborhood of any location on the traffic at the location in question. Cross correlation analysis is used to show the effect of the traffic from surrounding locations is significant, and has predictive power for forecasting a variable. A significant portion of this chapter appears in Chandra and Al-Deek (2008). For this, multivariate time series prediction techniques will have to be used. The next sections show the data used and the results of

cross correlation analysis on the data. Finally, the results and conclusions of the analysis are presented with pointers for future work.

### 4.1 Data Description

As was explained in the previous chapter, the relationship between any two variables is assessed by their covariance or correlation. The relationship between past values of a variable with its future values is assessed by autocorrelation. Cross correlation is used to assess the relationship between past values of one variable with future values of another. When the variables are time series, the stationarity of the series is critical. Non-stationarity could be present as a trend in the series or seasonality, which needs to be removed. A visual inspection of time series usually provides indication of non-stationarity. If there is a strong trend, then the autocorrelations decay very slowly. For seasonal trends, there are significant spikes at these seasonal periods. The removal of non-stationarity is achieved by differencing the time series till it is stationary. This ensures that the autocorrelations and the cross correlations die quickly and there are no long term trends in the variables.

In this chapter, the autocorrelations and cross correlations will be illustrated for speeds, volumes and occupancy at two sections – the downtown area of I-4 east bound (detector stations 33-37) and the area just west of Disney World on I-4 east bound (detector stations 6-10). The locations of these two sections are shown in Figure 4-1 and Figure 4-2. The data used is at 5-minute aggregated level. Station 35 and station 8 along with their upstream and downstream stations from the month of March, 2003 are used to illustrate these correlations.
Figure 4-1: Location of detector at station 35 and surrounding stations on I-4 eastbound

Figure 4-2: Location of detector at station 8 and surrounding stations on I-4 eastbound
4.2 Cross Correlation Analysis for Speeds and Volumes

The following figures, Figure 4-3 and Figure 4-4 show time series of speeds for a typical weekday in the close neighborhood of station 8 (1 mile upstream and 1 mile downstream) and station 35. It can be seen that the speeds in the section “move” together through time.

Figure 4-3: Speed profiles for a typical day in 1 mile radius around station 8
Figure 4-4: Speed profiles for a typical day in 1 mile radius around station 35

The level of the time series change throughout the day, and similar pattern repeats the next day. Therefore the speeds are expected to be non-stationary with slowly decaying autocorrelation. A look at the autocorrelations plots as given in Figure 4-5 confirms non-stationarity, due to the slow decay of the autocorrelations.
Figure 4-5: Auto correlations for speeds at detector station 35 and detector station 8 (shown up to 288 lags)

Differencing is usually used to deal with any apparent periodicity in the data. Often, the appropriate differencing scheme is found by trial and error, and the expected linear and periodic trends play a part in settling for the appropriate scheme. If the time series is increasing linearly (linear trend), a differencing at lag 1 removes linear trend. If the time series is cyclic, a seasonal differencing removes cyclic trends. A series of differencing schemes are tried to see which of them results in a stationary series. Since the data used here are 5-minute speeds, the relevant lags for differencing were –
Differencing at lag 1 – trend elimination

Differencing at lag 288 – daily periodicity elimination (24 hours*60 minutes / 5 minutes = 288 intervals)

Differencing at lag 2016 – weekly periodicity elimination(7 days * 24 hours* 60minutes / 5 minutes = 288 intervals)

Differencing at lags 1,288 – trend and daily periodic elimination

Differencing at lags 1, 2016 – trend and weekly periodic elimination

Figure 4-6 and Figure 4-7 show the results of these differencing strategies and the resulting autocorrelations for station 35 and station 8.

Figure 4-6: Auto correlations for differenced speed series for station 35
Figure 4-7: Auto correlations for differenced speed series for station 8
The autocorrelations decay faster when differencing is applied and therefore the series is deemed stationary. By differencing, instead of the actual speeds, the change in speeds with respect to the previous time period is considered for modeling and forecasting. Univariate time series models can be fit to these differenced time series at relevant orders that can be used for prediction. Similar behavior is observed at other locations. Therefore, speed series need to differenced atleast once at lag 1 to be stationary for computing cross-correlations.
Figure 4-8a) and Figure 4-8b) show the cross correlations of speeds from station 35 with one upstream (station 34) and one downstream station (station 36). Non-stationarity is evident from the cross correlation that dies very slowly, similar to autocorrelation. To make the series stationary, a differencing of 1 is applied.

**Figure 4-8 a):** Cross correlations of speeds from station 35 with upstream station 34

**Figure 4-8b):** Cross correlations of speeds from station 35 with upstream station 34

**Figure 4-8:** Cross correlations of speeds from station 35 with upstream and downstream stations.
If $s_{x,t}$ is the variable (speed) at station $x$ and time $t$ then by differencing at lag 1, the new series is

$$\nabla s_{x,t} = s_{x,t} - s_{x,t-1} = (1 - B)s_{x,t} = z_{x,t}$$

where $B(s_{x,t}) = s_{x,t-1}$ (B is a back-shift operator of $s_{x,t}$)

Similarly, if differencing is applied at lag $s$ then $\nabla^s s_{x,t} = s_{x,t} - s_{x,t-s} = (1 - B^s)s_{x,t}$ where $s = \text{lag } s$ (the period of seasonality) and $B^s(s_{x,t}) = s_{x,t-s}$. **Error! Reference source not found.**

shows a sample of short term cross correlation plot between differenced speeds (at lag 1) for upstream (Figure 4-9a) and downstream (Figure 4-9b) stations of station 35.

![Cross Correlations for differenced series Speed 35 (response) with Speed 34 (input)](image)

**Figure 4-9a:** Cross correlations between differenced speed series for station 35 and immediate upstream station 34.
Figure 4-9b: Cross correlations between differenced speed series for station 35 and immediate downstream station 36.

Figure 4-9: Cross correlations for differenced speed series for station 35 with immediate upstream and downstream stations.

An examination of Figure 4-9 for significant cross correlations reveals the following characteristics (as mentioned in section 3.2):

- In Figure 4-9a) $\rho_{z(t),z(t-h)} = \rho^-(h)$ for lags >0 (right quadrant). Here $\rho^-(h) > 0$ for $h = 2$, since there is a spike at lag 2 in the right quadrant, to the right of the vertical axis. Therefore, past values of speed at station 34 influence future values of 35.

- In Figure 4-9a) $\rho_{z(t),z(t+h)} = \rho^-(-h)$ for lag <0 (left quadrant). Here $\rho^-(-h) > 0$ for $|h|\leq 3$, since there are three spikes in the left quadrant, to the immediate left of the vertical axis. Therefore, past values of speed at station 35 influence future values of 34.
In Figure 4-9b) $\rho_{z(35,t),z(36,t-h)} = \rho^+(h)$ for lag >0. Here $\rho^+(h) > 0$ for $h <= 2$.

Therefore, past values of speed at station 36 influence future values of 35.

In Figure 4-9b) $\rho_{z(35,t),z(36,t+h)} = \rho^-(h)$ for lag <0. Here $\rho^-(h) > 0$ for $|h| <= 2$.

Therefore, past values of speed at station 35 influence future values of 36.

This shows that while past values at upstream locations influence the future values at downstream locations (station 34 with station 35, station 35 with station 36), past values at downstream also influence future values of the upstream locations (station 35 with station 34, station 36 with station 35). As we move further outward from station 35 in the upstream and downstream directions, as shown in Figure 4-10 the cross correlations persist, but are not as strong, and die down beyond station 32 in the upstream direction and 37 in the downstream direction.
Figure 4-10b) Cross correlations of differenced speeds between stations 1 mile downstream of station 35

Figure 4-10: Cross correlations of differenced speeds between stations 1 mile upstream and downstream of station 35

Similar results have been found in different sections on I-4 (see Figure 4-11 for cross correlations for differenced speeds at station 8 with its upstream and downstream neighbors).
Similar to the cross correlation functions shown in the above figures for differenced speeds between spatially separated locations, cross correlations can also be defined between other stationary variables (speed with volume, speed with occupancy, etc). These relationships shown graphically, prove that other variables carry predictive information in them.

Figure 4-12 shows the cross correlations of the differenced speeds at station 35 with differenced volumes from upstream (station 34) and downstream (station 36) stations. The cross correlations are not as strong as speed-speed relationships, but are still significant.
Figure 4-12a): Cross correlations for differenced speed series for station 35 with volumes from upstream station (34)

Figure 4-12b): Cross correlations for differenced speed series for station 35 with volumes from downstream station (36)

Figure 4-12: Cross correlations for differenced speed series for station 35 with volumes from upstream and downstream stations.
Figure 4-13 shows 5-minute volumes on a typical day for station 35 and its neighboring stations in a 1 mile radius. The figure illustrates that the volumes around surrounding stations also “follow” each other and therefore can be expected to be highly correlated across space and time, similar to the speed series. Autocorrelation plots show similar non-stationarities like speeds, and after differencing at lag 1, show a quick decline. Figure 4-14 shows the correlation of differenced volumes at station 35 with differenced volumes from upstream and downstream similar to Figure 4-9. From the steps described in interpreting Figure 4-9, it can be concluded that there are significant cross correlations for both the positive and negative lags. Therefore, it is expected that transfer function models will not be appropriate to capture this feedback process.

Figure 4-15 shows the cross correlations between the change in volumes (differenced volumes) at station 35 with the change in speeds at the upstream and downstream stations. It can be noticed in this figure too that the cross correlations at non-zero lags are higher which implies that there is more information likely to be extracted from past values at upstream and downstream locations than their current values, similar to Figure 4-9.
Figure 4-13: Volumes at stations in the neighborhood of station 35 and station 8
Figure 4-14a) Cross correlations for differenced volume series for station 35 with volumes from upstream station 34.

Figure 4-14b) Cross correlations for differenced volume series for station 35 with volumes from downstream station 36.

Figure 4-14: Cross correlations for differenced volume series for station 35 with volumes from upstream and downstream stations.
Figure 4-15a) Cross correlations for differenced volume series for station 35 with differenced speeds from upstream station 34

Figure 4-15b) Cross correlations for differenced volume series for station 35 with differenced speeds from downstream station 36

Figure 4-15: Cross correlations for differenced volume series for station 35 with differenced speeds from upstream and downstream stations.
Similar results have been found in the other sections on I-4 in Orlando, with volumes and occupancies, with little variations on the strength and significance of the cross correlations with distance.

The analysis of cross correlations is different from what has been reported in the literature so far. Cross correlations have been used to prove that the upstream locations can be used to improve predictions at the location in question. However, cross correlations at positive lags only have been used to advance the case of using upstream location data to predict the region in question. It has either been assumed that cross correlations are symmetric with respect to positive and negative lags or they have been completely neglected. Whenever cross correlations at both positive and negative lags are significant, feedback processes are involved and the effect of both upstream and downstream locations has to be considered simultaneously. In such cases, a univariate time-series methodology is inappropriate and a multivariate methodology is necessary.

However, the presence of cross correlations may not necessarily provide better information. The modeling of multivariate covariance induces complexities in model estimation, and forecasting. It has to be checked if modeling the cross correlation can provide better forecasts, in terms of various performance measures. In the presence of cross correlations, univariate time series can become incomplete specifications. This would handicap them of providing an analysis of the effects of shocks in the (correlated) variables to the variable in question. It can be argued that the time series of the speeds at the respective series are stationary but are not white noise (some autocorrelations persist at low lags). Therefore, the cross correlations between two locations can be confounded due to autocorrelations present at each location. To counter this argument, time series models are fit for each location and the resulting residuals from these series are then used.
to compute the cross correlations. The cross correlations at positive and negative lags are significant even after relevant univariate time series models have been fit to the respective stations.

Table 4-1 shows the ARIMA models fit to the stations 34, 35 and 36. Of the candidate models tested on the basis of Root Mean Square Error (RMSE), Mean Absolute Percent Error (MAPE) and Mean Absolute Error (MAE). RMSE, MAPE and MAE are defined as

\[ RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N}} \]
\[ MAPE = \frac{1}{N} \sum \left| \frac{\hat{y}_i - y_i}{y_i} \right| \times 100 \]
\[ MAE = \frac{1}{N} \sum \left| \hat{y}_i - y_i \right| \]

where \( y_i \) = observed value of the response
\( \hat{y}_i \) = forecasts for \( y_i \)
N = number of data points
Table 4-1: Univariate Time Series ARIMA Models Fit to Differenced Speeds from Station 34, Station 35, Station 36

<table>
<thead>
<tr>
<th>Station</th>
<th>Model</th>
<th>Model Details</th>
<th>Training Data</th>
<th>Test Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>MAPE</td>
</tr>
<tr>
<td>34</td>
<td>ARIMA</td>
<td>Subset ARIMA(0,1,7) q=1,5,6,7</td>
<td>4.26</td>
<td>6.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Subset ARIMA(0,1,2016) q=5,6,7,288,2016</td>
<td>4.25</td>
<td>6.4</td>
</tr>
<tr>
<td>34</td>
<td>ARIMA</td>
<td>ARIMA(0,1,q)X(0,1,1) q = 2,4,5,6,7 s=288</td>
<td>4.75</td>
<td>7</td>
</tr>
<tr>
<td>35</td>
<td>ARIMA</td>
<td>Subset ARIMA(0,1,7) q=2,3,6,7</td>
<td>4.05</td>
<td>6.22</td>
</tr>
<tr>
<td>35</td>
<td>ARIMA</td>
<td>Subset ARIMA(0,1,q) q=6,7,288,2016</td>
<td>4</td>
<td>6.25</td>
</tr>
<tr>
<td>35</td>
<td>ARIMA</td>
<td>ARIMA(0,1,q)X(0,1,1) q=1,6,7 s=288</td>
<td>4.3</td>
<td>6.9</td>
</tr>
<tr>
<td>36</td>
<td>ARIMA</td>
<td>ARIMA(0,1,2) q=1,2</td>
<td>3.88</td>
<td>6.5</td>
</tr>
<tr>
<td>36</td>
<td>ARIMA</td>
<td>Subset ARIMA(0,1,2016) q=1,2,288,2016</td>
<td>3.9</td>
<td>6.51</td>
</tr>
<tr>
<td>36</td>
<td>ARIMA</td>
<td>ARIMA(0,1,q)X(0,1,1) q=2,4,5,6,7 s=288</td>
<td>4.14</td>
<td>7.27</td>
</tr>
</tbody>
</table>

The residuals for ARIMA models do not show any auto correlation. However, it has been found that the cross correlations between residuals are significant.

The analysis of cross correlations is different from what has been reported in the literature so far. Cross correlations have been used to prove that the upstream locations can be used to improve predictions at the location in question. However, cross correlations at positive lags only have been used to advance the case of using upstream location data to predict the region in question. It has either been assumed that cross
correlations are symmetric with respect to positive and negative lags or they have been completely neglected. Whenever cross correlations at both positive and negative lags are significant, feedback processes are involved and the effect of both upstream and downstream locations has to be considered simultaneously. In such cases, a univariate time-series methodology is inappropriate and a multivariate methodology is necessary.

4.3 Conclusions

In this chapter, a case has been made to consider the correlations for traffic data (specifically, speeds and volumes) among different locations. For the purpose of this chapter, speeds and volumes from station 35 on the eastbound direction of I-4, Orlando are used, along with speeds and volumes from station 8. Cross correlation analysis is used to prove that there is significant relationship between the current value of speed at station in question and the past values of the speeds at both upstream as well as downstream stations. This is a significant result and contribution of this chapter and dissertation, as the current literature only models the effect of upstream locations on the current location.

Cross correlation analysis has also proved that the univariate prediction models fall short of modeling the impact of upstream and downstream conditions. Models that incorporate only upstream effects are therefore incomplete specifications, and multivariate models that handle feedback between input and response are required to better model the prediction of spatial time series.

The chapter has introduced multivariate cross correlation analysis to establish the predictive power of closely spaced neighbors (loop detector stations) in prediction of speeds on a freeway section. This analysis has been unique in the context of freeway
traffic prediction. The research introduced in this chapter would be further refined as described earlier to result in better prediction models for freeway traffic data.
CHAPTER 5. MULTIVARIATE PREDICTION OF FREEWAY TRAFFIC SPEEDS AND VOLUMES USING VECTOR AUTOREGRESSIVE MODELS²

The previous chapter had provided graphical representations of cross correlations among traffic variables (specifically speeds and volumes) across closely spaced loop detector stations (placed within 2 miles of each other). This chapter uses those relationships as basis to model traffic variables from 3 or 5 adjacent locations jointly using Vector Auto Regressive (VAR) models, for stations 35 and station 8. A significant portion of this chapter and the results belong to the paper accepted for publication in the Journal of Intelligent Transportation Systems (J-ITS).

Typically, the inputs to the traffic prediction system include the real-time and historical traffic data, location (or OD) information, and geometrics of the section, and incident characteristics if present. The simplest statistical approach is just to use the historical data based on the time of the day and use the measures of central tendency (mean, median or mode). This method is intuitive to the layman and is analogous to their perception of traffic conditions based on habit and experience. However, this method is crude and the errors from this method are usually higher than the more sophisticated techniques, specifically during peak times when the conditions are more dynamic than off peak times. The presence of incidents also complicates matters as the historical average method does not consider random events like incidents. The more sophisticated statistical approaches utilize the highly significant temporal dependencies among each of the traffic variables (flow, speed, etc) to predict the traffic variables. For closely spaced

² Chandra, R. S., and Al-Deek, H. *Comparison of Univariate and Multivariate Prediction of Freeway Traffic Speeds and Volumes Using Vector Autoregressive Models,* accepted for publication in the Journal of Intelligent Transportation Systems (J-ITS) for publication, December 2008
traffic data, correlations across space are also significant. The previous chapter has shown that inter-variable cross correlations are significant, and that they are bi-directional (in both positive and negative lags). This can be seen as a limitation of the traditional univariate time series models as the traffic data from neighboring sites can provide information that can be useful for prediction of traffic at any given site. Soft computing approaches like Artificial Neural Networks, Genetic Algorithms, etc., do not have the power of interpretability like the statistical models. These approaches are highly data centric and adapt themselves to the data supplied. As such, they require a long training period and it is difficult to isolate and interpret the effects of the dependent variables.

This chapter deals with estimating various Vector Auto Regressive models for three freeway traffic variables – speeds, volumes and occupancies. The estimated multivariate models are then used to forecast the respective variables and their performance is compared to baseline univariate models and other simple forecasts.

5.1 Data Description

The data used in this chapter is from the Interstate 4 (I-4) in Orlando, Florida. The 36 mile (80 km) section from Disney World area in the west and St.John’s River in the east, through downtown Orlando is instrumented with dual loop detectors spaced at approximately 0.5 miles (0.8 km). The traffic variables considered for prediction are the 5-minute aggregated speeds and 5 minute aggregated volumes in the eastbound direction on I-4 towards Daytona, as specified in the previous chapter. Traffic data is collected once every 30 seconds (for years 1999-2003), but it is also archived at 5 minute levels for year 2003. Two stations were selected to demonstrate the prediction approaches in this chapter, belonging to two sections of I-4. Station 8 (referred to as section 1) belongs to
the west (south) of I-4 in the Disney Area, Kissimmee, Station 35 (referred to as section 2) belongs to the downtown Orlando section. Each section consists of 2 stations upstream and 2 stations downstream to the specific station under study (total of 5 stations) for the purposes of demonstration. Data from all days from March 2003 is used in this chapter. Data from March 1st to March 24th are used for training and data from March 25th to March 31st are used for validation.

If \( s_{x,t} \) is the variable at station \( x \) and time \( t \) then by differencing at lag 1, the new series is
\[
\nabla s_{x,t} = s_{x,t} - s_{x,t-1} = (1 - L)s_{x,t} = z_{x,t}
\]
where \( L(s_{x,t}) = s_{x,t-1} \) (\( L \) is a function of \( s_{x,t} \)).

Similarly, differencing at lag \( k \) is given by
\[
\nabla s_{x,t} = s_{x,t} - s_{x,t-k} = (1 - L_k)s_{x,t} = y_{x,t}
\]

Simple ARIMA and Seasonal ARIMA (SARIMA) models were fit to the series differenced at specific lags to remove any trend or seasonality (refer chapter 4). These models were then compared to Vector Auto Regressive models that are described in the next section.

### 5.2 Vector Auto Regressive Model

The Vector Autoregressive (VAR) models exploit the dynamic interactions among interrelated time series. VAR models are extensively used in prediction of economic time series. When two interrelated variables are collected over time, it is reasonable to expect that the variables are correlated to the past lags of each other. Therefore, as data collection instruments collect interrelated data, VAR models can be used to predict the variables from past values of themselves and the variables collected along with them.
In other words, the VAR models are the multivariate extensions of the univariate AR models to the multivariate case. For the case of loop detectors, that typically report traffic flows, occupancies (single loops) and speeds (dual loops), the forecasting of flows, speeds and occupancies can use the past values of each other.

Formally, if we denote \( y_{1t}, y_{2t}, y_{3t}, \ldots y_{kt} \) as the \( k \) interrelated time series each of length \( t \), then we can denote the forecasting function as

\[
\hat{y}_{i,d+h} = f(y_{i,d}, y_{i,d-1}, y_{i,d-2}, \ldots) - \text{univariate case}
\]

\[
\hat{y}_{i,d+h} = f(y_{1d}, y_{1,d-1}, y_{1,d-2}, \ldots y_{2d}, y_{2,d-1}, y_{2,d-2}, \ldots y_{kd}, y_{k,d-1}, y_{k,d-2}, \ldots) - \text{multivariate case}
\]

where \( \hat{y}_{i,d+h} \) is the \( h \)-step forecast made at time \( t \) for variable \( i \)

\( y_{i,d} \) is the observed / measured value of variable \( i \) at time \( t \)

\( f(.) \) is the forecasting function

For the univariate case, a linear 1-step ahead (\( h=1 \)) forecasting function can be represented as

\[
\hat{y}_{i,d+1} = V_i + \phi_{11} y_{i,d} + \phi_{12} y_{i,d-1} + \cdots + \phi_{1,p_d} y_{i,d-p_d+1}
\]

which is the representative of the AR process of finite order (\( p \)) if the forecast errors are uncorrelated.

For a multivariate time series forecasting problem, the extension of the AR process to multivariate case gives rise to the representation

\[
\hat{y}_{i,d+1} = V_i + \phi_{11} y_{i,d} + \phi_{12} y_{i,d-1} + \cdots + \phi_{1,p_d} y_{1,d-p_d+1} + \phi_{21} y_{2,d} + \phi_{22} y_{2,d-1} + \cdots + \phi_{2,p_d} y_{2,d-p_d+1} + \cdots + \phi_{k1} y_{k,d} + \phi_{k2} y_{k,d-1} + \cdots + \phi_{kp_d} y_{k,d-p_d+1}
\]

where \( V_i \) is the intercept term
\( \phi_i, \phi_k \): coefficients of the AR and VAR processes respectively

If we denote \( y_t = (y_{1t}, y_{2t}, \ldots, y_{kt})' \), \( \hat{y}_t = (\hat{y}_{1t}, \hat{y}_{2t}, \ldots, \hat{y}_{kt})' \), \( v_1 = (v_{1t}, v_{2t}, \ldots, v_{kt})' \) and

\[
A_i = \begin{bmatrix}
\phi_{11,i} & \cdots & \phi_{1k,i} \\
\vdots & \ddots & \vdots \\
\phi_{1,i} & \cdots & \phi_{k,i}
\end{bmatrix}
\]

then the above equation can be generalized in the vector matrix form as

\[
\hat{y}_{t+1} = \nu + A_1 y_t + A_2 y_{t-1} + \ldots + A_p y_{t-p+1}
\]

This is the forecasting equation for the VAR given by

\[
y_t = \nu + A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_p y_{t-p} + u_t
\]

where \( u_t = (u_{1t}, u_{2t}, \ldots, u_{kt}) \) are the K-dimensional independently and identically distributed random vectors.

The above is a VAR model of order \( p \), and \( u_t \) is such that \( E(u_t) = 0 \), \( E(u_t u_t') = \Sigma_u \) and \( E(u_t u_s') = 0 \) for \( s \neq t \).

Moreover, this VAR(p) model is stable if

\[
\det(I_k - A_1 z - A_2 z^2 - \ldots - A_p z^p) \neq 0 \text{ for } |z| <= 1
\]

which is equivalent to having the roots of the characteristic polynomial outside the unit circle.

For traffic data arising from numerous detectors on a stretch of freeway, VAR models can be applied to identify and quantify the effect of different traffic variables at different time lags from the location itself as well as other locations. VAR models can be used to
describe and predict multiple spatially referenced time series as their formulation fits very well to the current study.

Let $y_t = (y_{1t}, ..., y_{kt})^\top$, $t = 0, 1, ...$ represent a $(kX1)$ vector of time series variables. Then a VAR processes of autoregressive order $p$ is written as

$$y_t = \nu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + u_t$$

Here $y_t$ denotes a $k$-dimensional time series vector which are observations from $k$ loop detectors under study at time $t$, $u_t$ is a $k$-dimensional vector of white noise process with $E(u_t) = 0$, $E(u_t u_s) = 0$ for all $t \neq s$ and $\phi_1, \phi_2, ..., \phi_p$ represent parameter matrices that have to be estimated. In another kind of model representation, a loop detector could measure $k$ variables (volume and occupancy for single loops; volume, occupancy and speed for dual loops). A similar kind of vector model representation could be used for forecasting purposes. Least Square or Maximum Likelihood methods are used to estimate the parameter matrices.

The influence of the upstream / downstream detectors can also be incorporated in the form of restrictions on the parameter matrices. Thus, to quantify the dependence of station $s$ to some upstream / downstream stations that influence the station in question, the row of the matrices corresponding to the station in question are constrained to be zeros for all columns other than the upstream / downstream stations. SAS / ETS software can be used to estimate VAR models.

VAR models assume stationarity, therefore the series are differenced and the VAR models of order $p$ are fit based on the minimum Schwarz Bayesian Criterion (SBC). The identification and estimation of VAR models is complicated, and it has been a common practice to estimate VAR models with maximum order $p$ and then minimize the
information criterion, by successively reducing the order $p$. Since differenced series are used, statistical tests are conducted to test for the presence of unit roots. If multivariate models are fit to integrated series (series that need differencing to be stationary) these series might not be stable as the autoregressive parameter is greater than 1. This implies that the series “explode” and will not move towards the series mean (Luktepohl, 1993). The Augmented Dickey Fuller test is used to check the null hypothesis of the presence of a unit root. The unit root test is based on the following equation where $y_t$ is a time series variable.

$$
\Delta y_t = \alpha y_{t-1} + \beta_1 \Delta y_{t-1} + \ldots + \beta_p \Delta y_{t-p} + \delta x_t + v_t
$$

The null hypothesis is that $H_0 : \alpha = 0$ and the alternate hypothesis is $H_a : \alpha < 0$. If we reject the null hypothesis, then the series $y_t$ is I(0). Since we perform this test with differenced speeds, if we reject the null hypothesis, then the speeds series is I(1), since the speeds are already differenced once.

## 5.3 Results

As was explained in the previous sections, a multivariate prediction approach is deemed to be more appropriate for the prediction of traffic series. Since each loop detector provides multivariate data (speeds, volumes, occupancies), at multiple locations, it is expected that the data is rich with multivariable correlations and cross correlations. The plots from the aforementioned sections prove that in addition to the information from the past of single variable at a single location, there is information available from other variables in surrounding locations that can be used for prediction.

In this chapter, the results of multivariate time series prediction of speeds and volumes for the representative stations at the three sections as mentioned earlier are
demonstrated. Though each loop detector at a location provides tri-variable data (flow, speed, occupancy), this chapter demonstrates the prediction of flow (volumes) and speeds only.

For the VAR formulation, speeds from station 35 are used as an instance. The same procedure is followed for station 8.

Let \( S_t = (S_{35,t}, S_{36,t}, S_{34,t}) \) represent the speeds at an instant \( t \) at stations 35, 36 (downstream to station 35) and 34 (upstream to station 35). For this 3-variable system, after \( S_t \) is suitably differenced so that it is stationary (usually \( d=1 \) for single time step change or \( d=288 \) for daily change), a VAR(p) model can be formulated as:

\[
\begin{bmatrix}
S_{35,t} \\
S_{34,t} \\
S_{36,t}
\end{bmatrix} = \begin{bmatrix}
\phi_{35,35} & \phi_{35,34} & \phi_{35,36} \\
\phi_{34,35} & \phi_{34,34} & \phi_{34,36} \\
\phi_{36,34} & \phi_{36,34} & \phi_{36,36}
\end{bmatrix} \begin{bmatrix}
S_{35,t-1} \\
S_{34,t-1} \\
S_{36,t-1}
\end{bmatrix} + \cdots + \begin{bmatrix}
\phi_{35,35} & \phi_{35,34} & \phi_{35,36} \\
\phi_{34,35} & \phi_{34,34} & \phi_{34,36} \\
\phi_{36,34} & \phi_{36,34} & \phi_{36,36}
\end{bmatrix} \begin{bmatrix}
S_{35,t-p} \\
S_{34,t-p} \\
S_{36,t-p}
\end{bmatrix}
\]

which can be reduced to univariate equation

\[
S_{35,t} = \phi_{35,35,t-1} * S_{35,t-1} + \phi_{35,34,t-1} * S_{34,t-1} + \phi_{35,36,t-1} * S_{36,t-1} \\
+ \phi_{35,35,t-2} * S_{35,t-2} + \phi_{35,34,t-2} * S_{34,t-2} + \phi_{35,36,t-2} * S_{36,t-2} + \cdots \\
+ \phi_{35,35,t-p} * S_{35,t-p} + \phi_{35,34,t-p} * S_{34,t-p} + \phi_{35,36,t-p} * S_{36,t-p}
\]

Similar univariate equations can be derived for \( S_{34,t} \) and \( S_{36,t} \). Since VAR models estimate three equations simultaneously, their complexity is comparable to estimation of three separate univariate equations. While VAR models might need more parameters to be estimated, the insignificant parameters are usually constrained to zero, so as to reduce to a manageable and intuitive model. Moreover, in addition to the improvement accuracy, VAR models provide a framework where the impact of change in one variable can be traced through all the variables (impulse response functions). This kind of a framework allows the analyst to answer questions like “how long and how much would
the impact of reduction of speed at station x exist on its upstream / downstream stations?” Such an analysis would not be possible through simple univariate formulations. These advantages make the VAR models attractive for real-time traffic prediction and forecasting purposes, when faced with unexpected incidents (or shocks).

The order (p) selection is done by selecting $p$ that minimizes the SBC. Table 5-1 shows the comparison of different univariate and multivariate prediction mechanisms in terms of two measures of performance for training dataset and validation data as explained in the data description. These measures are the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). These are defined as

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}}$$

and

$$MAPE = \frac{\sum|y_i - \hat{y}_i|}{n}$$
Table 5-1: Comparison of the Univariate Time Series ARIMA Models and Multivariate VAR models Fit to Differenced Speeds from Station 8, Station 35

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model</th>
<th>Input Variables*</th>
<th>Training Data</th>
<th>Test Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>MAPE</td>
</tr>
<tr>
<td>8 Speed</td>
<td>Historical Average</td>
<td>None</td>
<td>7.06</td>
<td>5.12</td>
</tr>
<tr>
<td>8 Speed</td>
<td>SARIMA(1,0,1)x(0,1,1) s=2016</td>
<td>None</td>
<td>2.96</td>
<td>3.89</td>
</tr>
<tr>
<td>8 Speed</td>
<td>ARIMA(0,1,2)</td>
<td>None</td>
<td>2.57</td>
<td>2.98</td>
</tr>
<tr>
<td>8 Speed</td>
<td>SARIMA(0,1,2)x(0,1,1) s=288</td>
<td>None</td>
<td>2.69</td>
<td>3.19</td>
</tr>
<tr>
<td>8 Speed</td>
<td>VAR(5)</td>
<td>1 up speed, 1 down speed</td>
<td>1.54</td>
<td>1.97</td>
</tr>
<tr>
<td>8 Speed</td>
<td>VAR(7)</td>
<td>2 up speed, 2 down speed</td>
<td>1.52</td>
<td>1.97</td>
</tr>
<tr>
<td>8 Speed</td>
<td>VAR(6)</td>
<td>1 up volume, 1 down volume</td>
<td>1.57</td>
<td>1.98</td>
</tr>
<tr>
<td>8 Speed</td>
<td>VAR(8)</td>
<td>1 up volume, 1 down speed</td>
<td>1.54</td>
<td>1.98</td>
</tr>
<tr>
<td>8 Speed</td>
<td>VAR(5)</td>
<td>1 up speed, 1 down volume</td>
<td>1.54</td>
<td>1.97</td>
</tr>
<tr>
<td>35 Speed</td>
<td>Historical Average</td>
<td>None</td>
<td>12.01</td>
<td>18.08</td>
</tr>
<tr>
<td>35 Speed</td>
<td>SARIMA(1,0,1)x(0,1,1) s=2016</td>
<td>None</td>
<td>4.89</td>
<td>6.38</td>
</tr>
<tr>
<td>35 Speed</td>
<td>Subset ARIMA(0,1,7)</td>
<td>None</td>
<td>4.05</td>
<td>6.22</td>
</tr>
<tr>
<td>35 Speed</td>
<td>SARIMA(0,1,q)x(0,1,1) s= 288</td>
<td>None</td>
<td>4.3</td>
<td>6.9</td>
</tr>
<tr>
<td>35 Speed</td>
<td>VAR(4)</td>
<td>1 up speed, 1 down speed</td>
<td>3.16</td>
<td>5.86</td>
</tr>
<tr>
<td>35 Speed</td>
<td>VAR(6)</td>
<td>2 up speed, 2 down speed</td>
<td>3.13</td>
<td>5.36</td>
</tr>
<tr>
<td>35 Speed</td>
<td>VAR(5)</td>
<td>1 up volume, 1 down volume</td>
<td>3.2</td>
<td>6.12</td>
</tr>
<tr>
<td>35 Speed</td>
<td>VAR(4)</td>
<td>1 up volume, 1 down speed</td>
<td>3.18</td>
<td>6.09</td>
</tr>
<tr>
<td>35 Speed</td>
<td>VAR(4)</td>
<td>1 up speed, 1 down volume</td>
<td>3.19</td>
<td>6.25</td>
</tr>
</tbody>
</table>

* For input variables, up refers to upstream variable, down refers to downstream variable

Table 2 shows the parameters of the VAR model estimated for the multi variable system of speeds from station 33, station 34, station 35, station 36 and station 37.
Table 5-2: VAR Models Fit to Differenced Speeds from Station 33, Station 34, Station 35, Station 36, Station 37

Table 5-2a) Parameters for a VAR(4) model from Station 34, Station 35, Station 36

<table>
<thead>
<tr>
<th>Order of Difference</th>
<th>Response Station</th>
<th>Input Station</th>
<th>Lag</th>
<th>Relative location to Response station</th>
<th>AR Coefficient</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Speed 34</td>
<td>Speed 34</td>
<td>1</td>
<td>Current</td>
<td>-0.05417</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 34</td>
<td>Speed 35</td>
<td>1</td>
<td>Downstream</td>
<td>0.21407</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 34</td>
<td>Speed 36</td>
<td>1</td>
<td>Downstream</td>
<td>0.12082</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 34</td>
<td>Speed 34</td>
<td>2</td>
<td>Current</td>
<td>-0.05684</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 34</td>
<td>Speed 35</td>
<td>2</td>
<td>Downstream</td>
<td>0.10531</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 34</td>
<td>Speed 36</td>
<td>2</td>
<td>Downstream</td>
<td>0.05291</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 34</td>
<td>Speed 34</td>
<td>3</td>
<td>Current</td>
<td>-0.04263</td>
<td>0.0007</td>
</tr>
<tr>
<td>1</td>
<td>Speed 34</td>
<td>Speed 35</td>
<td>3</td>
<td>Downstream</td>
<td>0.07859</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 34</td>
<td>Speed 36</td>
<td>3</td>
<td>Downstream</td>
<td>0.0278</td>
<td>0.0292</td>
</tr>
<tr>
<td>1</td>
<td>Speed 34</td>
<td>Speed 34</td>
<td>4</td>
<td>Current</td>
<td>-0.05822</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 34</td>
<td>Speed 35</td>
<td>4</td>
<td>Downstream</td>
<td>0.05569</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 35</td>
<td>Speed 35</td>
<td>1</td>
<td>Current</td>
<td>-0.0721</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 35</td>
<td>Speed 36</td>
<td>1</td>
<td>Downstream</td>
<td>0.19266</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 35</td>
<td>Speed 34</td>
<td>2</td>
<td>Upstream</td>
<td>0.14669</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 35</td>
<td>Speed 35</td>
<td>2</td>
<td>Current</td>
<td>-0.08399</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 35</td>
<td>Speed 36</td>
<td>2</td>
<td>Downstream</td>
<td>0.11746</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 35</td>
<td>Speed 34</td>
<td>3</td>
<td>Upstream</td>
<td>0.07652</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 35</td>
<td>Speed 35</td>
<td>3</td>
<td>Current</td>
<td>-0.07839</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 35</td>
<td>Speed 36</td>
<td>3</td>
<td>Downstream</td>
<td>0.06751</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 35</td>
<td>Speed 34</td>
<td>4</td>
<td>Upstream</td>
<td>0.03119</td>
<td>0.003</td>
</tr>
<tr>
<td>1</td>
<td>Speed 35</td>
<td>Speed 35</td>
<td>4</td>
<td>Current</td>
<td>-0.05621</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 36</td>
<td>Speed 34</td>
<td>1</td>
<td>Upstream</td>
<td>-0.06902</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 36</td>
<td>Speed 35</td>
<td>1</td>
<td>Upstream</td>
<td>0.24643</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 36</td>
<td>Speed 36</td>
<td>1</td>
<td>Current</td>
<td>-0.2095</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 36</td>
<td>Speed 35</td>
<td>2</td>
<td>Upstream</td>
<td>0.25351</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 36</td>
<td>Speed 36</td>
<td>2</td>
<td>Current</td>
<td>-0.17786</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 36</td>
<td>Speed 35</td>
<td>3</td>
<td>Upstream</td>
<td>0.12818</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 36</td>
<td>Speed 36</td>
<td>3</td>
<td>Current</td>
<td>-0.11967</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 36</td>
<td>Speed 35</td>
<td>4</td>
<td>Upstream</td>
<td>0.11296</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>Speed 36</td>
<td>Speed 36</td>
<td>4</td>
<td>Current</td>
<td>-0.08584</td>
<td>0.0001</td>
</tr>
</tbody>
</table>
The model for the differenced speeds for station 35 based on Table 5-2a) can now be written as:

\[
S_{35,t} = S_{35,t-1} - 0.0721*(S_{35,t-1} - S_{35,t-2}) + 0.1466*(S_{34,t-2} - S_{34,t-3}) + 0.1926*(S_{36,t-1}-S_{36,t-2}) + 0.117 *(S_{36,t-2} - S_{36,t-3}) -0.0839*(S_{35,t-2} - S_{35,t-3}) + 0.076*(S_{34,t-3}-S_{34,t-4}) - 0.078*(S_{35,t-3}-S_{35,t-4}) + 0.06*( S_{36,t-3} -S_{36,t-4}) + 0.03*( S_{34,t-4} -S_{34,t-5}) - 0.05*( S_{35,t-4} -S_{35,t-5})
\]

The residuals from these models are expected to be devoid of any further cross correlations within the extent to which variables from neighboring locations have been added. As is evident from these parameters, the past values at upstream and downstream stations are significant. An examination of the cross correlations between the residuals of
speeds at station 35 and its surrounding stations (Figure 5-1) illustrates the success of VAR in accounting for cross correlations.

Figure 5-1b) Cross correlations of residuals from VAR(4) of differenced speeds from Station 35 (RES1) with residuals from Station 34(RES2).

Figure 5-1b) Cross correlations of residuals from VAR(4) of differenced speeds from Station 35 (RES1) with residuals from Station 36(RES3).

Figure 5-1: Cross correlations of residuals from VAR(4) models on Station 35 with Station 34 and Station 36.
The performance of VAR models for volumes vis-à-vis univariate models is provided in Table 5-3, with similar results.

**Table 5-3: Comparison of the Univariate Time Series ARIMA Models and Multivariate VAR models Fit to Differenced Volumes from Station 8, Station 35**

<table>
<thead>
<tr>
<th>Station</th>
<th>Model</th>
<th>Input Variables*</th>
<th>Training Data</th>
<th>Test Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>MAPE</td>
</tr>
<tr>
<td>8 Volume</td>
<td>Historical Average</td>
<td>None</td>
<td>33.56</td>
<td>16.27</td>
</tr>
<tr>
<td>8 Volume</td>
<td>SARIMA(1,0,1)*(0,1,1) s=2016</td>
<td>None</td>
<td>26.53</td>
<td>12.34</td>
</tr>
<tr>
<td>8 Volume</td>
<td>ARIMA(3,1,4)*(1,0,0) s = 288</td>
<td>None</td>
<td>25.01</td>
<td>11.31</td>
</tr>
<tr>
<td>8 Volume</td>
<td>SARIMA(2,1,3)*(0,1,1) s = 288</td>
<td>None</td>
<td>25.49</td>
<td>11.87</td>
</tr>
<tr>
<td>8 Volume</td>
<td>VAR(8)</td>
<td>1 up, 1 down</td>
<td>22.72</td>
<td>9.31</td>
</tr>
<tr>
<td>8 Volume</td>
<td>VAR(8)</td>
<td>2 up, 2 down</td>
<td>21.5</td>
<td>9.22</td>
</tr>
<tr>
<td>8 Volume</td>
<td>VAR(7)</td>
<td>1 up speed, 1 down speed</td>
<td>23.38</td>
<td>9.71</td>
</tr>
<tr>
<td>8 Volume</td>
<td>VAR(7)</td>
<td>1 up volume, 1 down speed</td>
<td>22.88</td>
<td>9.37</td>
</tr>
<tr>
<td>8 Volume</td>
<td>VAR(7)</td>
<td>1 up speed, 1 down volume</td>
<td>23.3</td>
<td>9.61</td>
</tr>
<tr>
<td>35 Volume</td>
<td>Historical Average</td>
<td>None</td>
<td>51.34</td>
<td>19.15</td>
</tr>
<tr>
<td>35 Volume</td>
<td>SARIMA(1,0,1)*(0,1,1) s=2016</td>
<td>None</td>
<td>33.52</td>
<td>12.53</td>
</tr>
<tr>
<td>35 Volume</td>
<td>ARIMA(0,1,1)X(1,0,0) s = 288</td>
<td>None</td>
<td>29.26</td>
<td>10.84</td>
</tr>
<tr>
<td>35 Volume</td>
<td>SARIMA(0,1,3)X(0,1,1) s=288</td>
<td>None</td>
<td>29.41</td>
<td>10.36</td>
</tr>
<tr>
<td>35 Volume</td>
<td>VAR(4)</td>
<td>1 up volume, 1 down volume</td>
<td>27.72</td>
<td>9.58</td>
</tr>
<tr>
<td>35 Volume</td>
<td>VAR(6)</td>
<td>2 up volume, 2 down volume</td>
<td>26.66</td>
<td>9.46</td>
</tr>
<tr>
<td>35 Volume</td>
<td>VAR(5)</td>
<td>1 up speed, 1 down speed</td>
<td>27.79</td>
<td>9.73</td>
</tr>
<tr>
<td>35 Volume</td>
<td>VAR(4)</td>
<td>1 up volume, 1 down speed</td>
<td>27.69</td>
<td>9.64</td>
</tr>
<tr>
<td>35 Volume</td>
<td>VAR(4)</td>
<td>1 up speed, 1 down volume</td>
<td>27.89</td>
<td>9.66</td>
</tr>
</tbody>
</table>

* For input variables, up refers to upstream variable, down refers to downstream variable

It can be seen that the VAR model for the volumes outperform the ARIMA models. Also, it is interesting to note that for volumes, the model with 1 upstream station volume and 1 downstream station speed performed almost as good as the model with 2
upstream and 2 downstream volumes. However, the speeds do not appear to be very sensitive to this combination of data. Table 5-4 shows the result of the Augmented Dickey Fuller test for station 8 and its neighbors and station 35 and its neighbors for speeds. These tests confirm that the differenced speed series considered do not have unit roots in their AR parts.

Table 5-4: Augmented Dickey Fuller Statistics for Testing Unit Roots in Differenced Speeds and Volumes at Station 35 and Station 8 and their Upstream and Downstream Stations

<table>
<thead>
<tr>
<th>Variable</th>
<th>Augmented Dickey-Fuller test statistic</th>
<th>p-value</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speeds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D(speed 8)</td>
<td>-36.98777</td>
<td>0</td>
<td>Reject Null Hypothesis</td>
</tr>
<tr>
<td>D(speed 9)</td>
<td>-38.09061</td>
<td>0</td>
<td>Reject Null Hypothesis</td>
</tr>
<tr>
<td>D(speed 7)</td>
<td>-46.08475</td>
<td>0.0001</td>
<td>Reject Null Hypothesis</td>
</tr>
<tr>
<td>D(speed 35)</td>
<td>-90.75194</td>
<td>0.0001</td>
<td>Reject Null Hypothesis</td>
</tr>
<tr>
<td>D(speed 36)</td>
<td>-71.79453</td>
<td>0.0001</td>
<td>Reject Null Hypothesis</td>
</tr>
<tr>
<td>D(speed 34)</td>
<td>-88.73741</td>
<td>0.0001</td>
<td>Reject Null Hypothesis</td>
</tr>
<tr>
<td>Volumes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D(volume 8)</td>
<td>-17.28984</td>
<td>0</td>
<td>Reject Null Hypothesis</td>
</tr>
<tr>
<td>D(volume 7)</td>
<td>-16.29281</td>
<td>0</td>
<td>Reject Null Hypothesis</td>
</tr>
<tr>
<td>D(volume 6)</td>
<td>-16.31051</td>
<td>0</td>
<td>Reject Null Hypothesis</td>
</tr>
<tr>
<td>D(volume 35)</td>
<td>-68.65729</td>
<td>0.0001</td>
<td>Reject Null Hypothesis</td>
</tr>
<tr>
<td>D(volume 36)</td>
<td>-53.08341</td>
<td>0.0001</td>
<td>Reject Null Hypothesis</td>
</tr>
<tr>
<td>D(volume 34)</td>
<td>-65.62205</td>
<td>0.0001</td>
<td>Reject Null Hypothesis</td>
</tr>
</tbody>
</table>

Table 5-5 shows the effect of including upstream and downstream stations that are farther than the immediate neighbors. For example, station 8’s immediate neighbors are station 7 and 9, each at a distance of approximately half a mile from station 8. Station 6 and station 10 are approximately 1 mile from station 8. Station 5 and station 11 are 1.5 miles from station 8. The VAR model for speeds from station 8 (and similarly for station 35) from neighbors at different distances is shown in Table 5-5. The RMSE and MAPE for these models can be seen to increase slightly over increasing distances from neighbors. The cross correlation function also shows weaker correlations at low lags as
distance increases. However, issue of distance at which a neighbor can be deemed not useful for prediction is non-trivial and should be explored in future research.

Table 5-5: Effect of Distance on Measures of Performance of VAR Models for Speeds from Station 8 and Station 35

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Model</th>
<th>Input Stations*</th>
<th>Spacing between consecutive stations</th>
<th>Training Data</th>
<th>Test Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 Speed</td>
<td>VAR(5)</td>
<td>7, 9</td>
<td>0.5 miles</td>
<td>1.54</td>
<td>1.85</td>
</tr>
<tr>
<td>8 Speed</td>
<td>VAR(7)</td>
<td>6,7,9,10</td>
<td>0.5 miles</td>
<td>1.52</td>
<td>1.84</td>
</tr>
<tr>
<td>8 Speed</td>
<td>VAR(5)</td>
<td>6,10</td>
<td>1 mile</td>
<td>1.68</td>
<td>2.12</td>
</tr>
<tr>
<td>8 Speed</td>
<td>VAR(8)</td>
<td>5, 11</td>
<td>1.5 miles</td>
<td>1.7</td>
<td>2.19</td>
</tr>
<tr>
<td>35 Speed</td>
<td>VAR(4)</td>
<td>34, 36</td>
<td>0.5 miles</td>
<td>3.16</td>
<td>3.2</td>
</tr>
<tr>
<td>35 Speed</td>
<td>VAR(6)</td>
<td>33, 34, 36, 37</td>
<td>0.5 miles</td>
<td>3.13</td>
<td>3.22</td>
</tr>
<tr>
<td>35 Speed</td>
<td>VAR(7)</td>
<td>33,37</td>
<td>1 mile</td>
<td>3.57</td>
<td>3.79</td>
</tr>
<tr>
<td>35 Speed</td>
<td>VAR(8)</td>
<td>32,38</td>
<td>1.5 miles</td>
<td>3.64</td>
<td>3.85</td>
</tr>
</tbody>
</table>

*Endogenous variables from input stations are speeds

In this chapter, the effect of occupancies on prediction has not been tested. In the wake of the observation that volumes might be sensitive to upstream volumes and downstream speeds, it might be interesting to investigate VAR models with occupancies as the included (endogenous) variables.

5.4 Conclusions

In this chapter, multivariate time series models were defined in the context of loop detector data. Different locations (closely spaced loop detectors) were considered as different variables that were jointly correlated, which needed to be estimated. For the purpose of this chapter, speeds and volumes from station 8 and station 35 were used as response variables. The multivariate specification ensures that all endogenous variables will be estimated jointly, and this is comparable in complexity to estimating as many different univariate models, with the additional constraint due to covariance. Also, both
upstream and downstream locations are used as endogenous variables in this specification. This is justified due to the close locations of the loop detectors, due to which feedback effects are significant. This is a significant result and contribution of this chapter, as the current literature only models the effect of upstream locations on the current location. Models that incorporate only upstream effects are therefore incomplete specifications, and multivariate models that handle feedback between input and response are required to better model the prediction of spatial time series.

VAR model analysis was performed to show that VAR models are comparatively better than ARIMA models for prediction. It has been shown that VAR models have lower errors than ARIMA models on both the training and test data sets. The residuals from the VAR models do not have significant cross correlations – this implies that no further information can be extracted from the neighboring upstream and downstream stations for prediction.

Further investigation of VAR models is required. Stationarity and constant variance is an assumption required in the formulation of the VAR model, and it is possible that these assumptions are not realistic. The constraints on the VAR model to avoid problems of over parameterization also need to be investigated. Moreover, the application of a single VAR model for the whole 50 mile instrumented section of I-4 can be cumbersome and impractical. It is possible that there are spatial clusters – groups of stations that behave in a similar way. For example, the Disney Area (station 8) would have a different traffic behavior than the Downtown section (station 35). The rush hour traffic behavior would be different from the off-peak traffic behavior that would warrant different VAR models for different time periods. In other words, further research is
prompted that can answer how many upstream and downstream locations are required to obtain the best possible prediction of traffic condition at the location in question.

The chapter has introduced VAR models as a forecasting model for prediction of speeds and volumes on a freeway section. Such analysis has been unique in the context of freeway traffic prediction. On the other hand, it also needs to be investigated how the performance of the VAR model changes with respect to different temporal aggregation levels, as traffic prediction is carried out at different aggregation levels. The next chapter focuses on this issue and answers the question about the aggregation level of data for which, the application of VAR models are appropriate.
CHAPTER 6. MULTIVARIATE MODELS FOR FREEWAY TRAFFIC PREDICTION AT THREE TIME SCALES ON INTERSTATE 4

6.1 Introduction

The previous chapters had elaborated on the nature of cross correlations between traffic variables that are collected by sensors separated by short distances (less than 2 miles), and ways to estimate the multivariate forecasting functions. These models were specified on speeds and volumes collected at 5-minute intervals and their one-step ahead forecasts were compared. VAR models performed better than univariate models and historical averages. Short term traffic forecasting, however, is the prediction of traffic variables up to future periods – typically for intervals less than an hour. To this effect, different prediction algorithms have been applied at different data aggregation levels.

The statistical, non-parametric and artificial intelligence approaches to traffic prediction have had different rates success on different data. While data that is very noisy is difficult to be modeled, there is utility for data that is collected at low aggregation intervals. If the data is too smoothed out, important characteristics and variations might be lost to the analyst and the whole traffic operations might suffer. Therefore, Traffic Management centers tend to maintain data at different levels of aggregation, in such a way, that the data can be used for a variety of purposes. Raw freeway data is typically collected at 20s – 60s frequency, but may be archived at a higher aggregation level, for e.g., 5 minutes, 10 minutes, 15 minutes or higher. Different agencies and stake holders

---

3 Chandra, R. S., and AL-DEEK, H. *Comparison of Univariate and Multivariate Models for Freeway Traffic Prediction at Three Time Scales on Interstate 4,* Submitted to the Transportation Engineering Journal of ASCE for publication, Originally submitted in December 2008, completed one round of revisions and went into the second cycle of review, March 2009
require different details. Therefore, for any forecasting system, it is essential to test the robustness of its performance against different level of detail of the variables.

The performance of multivariate forecasting of traffic variables between a location and its immediate neighborhood are discussed in this chapter. This performance is described in terms of correlations and prediction power at various scales of time (or resolution where the data is aggregated). The next section reviews cross correlation analysis very briefly. The description of multivariate traffic relationships at various time scales and the performance of multivariate time series correlation techniques at these time scales are described in the following sections. The next sections show the data used and the results of multivariate regression models on the data. Finally, the conclusions of the analysis are presented with pointers for future work.

6.2 Data Description

The data used in this chapter is from the Interstate 4 (I-4) in Orlando, Florida, from the same locations used in the previous chapters. In addition to, station 8 and station 35 in the east bound direction on I-4, the analysis is now includes data from station 60. These three stations are randomly selected from the eastern, central and western instrumented portions of I-4 in Central Florida. The traffic variables considered are the volumes aggregated at 5-minute, 10-minute and 15-minute levels. Data from all days from March 2003 is used in this chapter.

6.3 Methodology

In the literature, cross-correlation among multivariate traffic series has not received extensive attention, with the exception of Stathopoulous (2001), Yue (2006) and Chandra
and Al-Deek (2008). The relationship between the (time) lagged relationships between traffic variables collected at the same location as well as variables collected across multiple locations would be instructive of the multivariate relationships among data. These cross correlation analyses were performed between different series, after removing any non-stationarity.

The volume series at each station was aggregated at 5 minutes, 10 minutes and 15 minutes, but the cross correlations for 5 minute and 15 minute aggregated volumes for station 35 with its upstream and downstream station from the month of March, 2003 are shown for demonstration. If \( V_{x,t} \) is the variable (flow) at station \( x \) and time \( t \) then by differencing at lag 1, the new series is

\[
\nabla V_{x,t} = V_{x,t} - V_{x,t-1} = (1 - L)V_{x,t} = v_{x,t}
\]

where \( L(V_{x,t}) = v_{x,t-1} \) (L is called a “lag function” of \( V_{x,t} \)).

Simple differencing at lag 1 removes any non-stationarity due to time trend in the mean of the variable. Similarly, if there is non-stationarity due to periodic / seasonal trends, then a seasonal differencing is applied. This can be shown as

\[
\nabla^s V_{x,t} = V_{x,t} - V_{x,t-s} = (1 - B^s)V_{x,t} = v_{x,t}^s
\]

where \( s \) = seasonal period.

For 5-minute aggregated data, the daily seasonal period is 288 5-minute intervals, while for 15-minute aggregated data, daily seasonal period is 96 15-minute intervals. The autocorrelations for each individual series provide an indication of the appropriate univariate ARIMA time series models that can be used. Cross correlation analysis was performed with the different elements of traffic state – volumes, speeds and occupancies,
collected from the same location, and from surrounding locations. Cross correlation for
two series $Z_t$ and $X_t$ at lag $h$ is defined as

$$
\rho(z_t, x_{t-h}) = \rho_{zt}(h) = \frac{\sum_{i=1}^{n-h} (z_i - \bar{z})(x_{i-h} - \bar{x})}{\sqrt{\sum_{i} (z_i - \bar{z})^2} \sqrt{\sum_{i} (x_{i-h} - \bar{x})^2}} = \frac{\text{cov}(Z_t,X_{t-h})}{s_z s_x}
$$

Cross correlation function can be used to determine time lagged speed-volume-
occupancy relationships. This has the potential to answer the question about the utility of
past speed / occupancy values to predict future values of volumes (or other traffic
variables).

In the case of significant cross correlations at positive and negative lags, Vector
Auto Regressive (VAR) models are considered more appropriate to model the feedback
relationship among the variables.

6.3.1 Vector Auto Regressive Model

The Vector Autoregressive (VAR) models exploit the dynamic interactions among
interrelated time series. VAR models are extensively used in prediction of economic
time series. When multiple interrelated variables are collected over time, it is reasonable
to expect that these variables are correlated to the past values of one another. Therefore,
as data collection instruments collect interrelated data, VAR models can be used to
predict the variables from past values of themselves and the variables collected along
with them.

Formally, if we denote $y_{1t}$, $y_{2t}$, $y_{3t}$, … $y_{kt}$ as the $k$ interrelated time series each of
length $t$, then we can denote the forecasting function as

$$
y_{i,d+h} = \hat{f}(y_{i,d}, y_{i,d-1}, y_{i,d-2}, \ldots) \text{ - univariate case}
$$
\[ \hat{y}_{i,t+h} = f(y_{i,t}, y_{i,t-1}, y_{i,t-2}, \ldots, y_{k,t}, y_{k,t-1}, y_{k,t-2}, \ldots) \] - multivariate case

where \( \hat{y}_{i,t+h} \) is the h-step forecast made at time t for variable i

\[ y_{i,t} \] is the observed / measured value of variable i at time t

\( f(.) \) is the forecasting function

For a multivariate time series, the forecasting equation can be represented as

\[ \hat{y}_{i,t+1} = \beta_{i1} + \phi_{i1,1} y_{i,t} + \phi_{i1,2} y_{i,t-1} + \ldots + \phi_{i1,p} y_{i,t-p+1} + \phi_{i2,1} y_{2,t} + \phi_{i2,2} y_{2,t-1} + \ldots + \phi_{i2,p} y_{2,t-p+1} + \ldots + \phi_{ikp,1} y_{k,t} + \phi_{ikp,2} y_{k,t-1} + \ldots + \phi_{ikp,p} y_{k,t-p+1} \]

where \( \beta_{i} \) is the intercept term

The above is a VAR model of order p, and \( u_{t} \) is such that \( E(u_{t}) = 0 \), \( E(u_{t}u_{t}') = \Sigma_{u} \) and \( E(u_{t}u_{t}') = 0 \) for \( s \neq t \). These are multivariate versions of zero mean \((E(\epsilon_{t}) = 0)\), constant variance \((E(\epsilon_{t}^{2}) = \sigma^{2})\), and zero covariance \((E(\epsilon_{t}\epsilon_{s}) = 0 \text{ for } t \neq s)\) conditions of random errors \( \epsilon_{t} \) required for univariate time series model.

A VAR(p) is stable if

\[ \det(I_{k} - A_{1}z-A_{2}z^{2}-\ldots-A_{p}z^{p}) \neq 0 \text{ for } |z| \leq 1 \]

where \( I_{k} \) is the identity matrix,

\( A_{1}, A_{2}, \ldots A_{p} \) are the coefficient matrices

\( z \) is the variable in the polynomial

\( \det() \) is the determinant function

which is equivalent to having the roots of the characteristic polynomial outside the unit circle.

The stability of VAR models has an important implication in their interpretation.

Specifically, a stable VAR model can be alternatively represented as a linear combination
of past errors – this is known as the Moving Average representation analogous to univariate time series models. Such a model is known as a “causal model” and can be used to trace the effects of shocks in one of the variables throughout the system. For traffic data arising from numerous detectors on a stretch of freeway, causal VAR models can be applied to identify and quantify the effect of different traffic variables at different time lags from the location itself as well as other locations. VAR models can be used to describe and predict multiple spatially referenced time series as their formulation fits very well to the current study.

The influence of the upstream / downstream detectors can also be incorporated in the form of restrictions on the parameter matrices. Thus, to quantify the dependence of station $s$ to some upstream / downstream stations that influence the station in question, the row of the matrices corresponding to the station in question are constrained to be zeros for all columns other than the upstream / downstream stations. Alternatively, statistical significance of parameters can be used to constrain the insignificant parameters to zero. SAS / ETS software can be used to estimate VAR models.

VAR models assume stationarity, therefore the series are differenced and the VAR models of order $p$ are fit based on the minimum Schwarz Bayesian Criterion (SBC) (Lutkepohl, 1993).

6.4 Results

The auto-correlations of the differenced volume series for station 35 at 5-minute aggregation and 15-minute aggregation levels were computed and shown in Figure 6-1a and Figure 6-1b.
Figure 6-1a-1: ACF of volume series at station 35 differenced at lag=1

Figure 6-1a-2: ACF of volume series at station 35 differenced at lag=288

Figure 6-1a-3: ACF of volume series at station 35 differenced at lag=2016
Figure 6-1a-4 ACF of volume series at station 35 differenced at lag=288 and lag=1

Spikes at low lags around 0
Spikes at lags around 288 (daily lags)

Figure 6-1a-5. ACF of volume series at station 35 differenced at lag=2016 and lag=1

Spikes at low lags around 0
Spikes at lags around 2016 (weekly lags)

Figure 6-1a. Auto correlations for differenced volume at station 35 at 5-minute aggregation.
Figure 6-1b 1. Auto correlation for differenced volume at station 35 at lag=1

Figure 6-1b 2. Auto correlation for differenced volume at station 35 at lag=96

Figure 6-1b 3. Auto correlation for differenced volume at station 35 at lag=672
Figure 6-1b 4. Auto correlation for differenced volume at station 35 at lag=96 and lag=1

Figure 6-1b5. Auto correlation for differenced volume at station 35 at lag=672 and lag=1

Figure 6-1b. Auto correlations for differenced volume at station 35 at 15-minute aggregation.

Figure 6-1: Auto correlations for differenced volume series for station 35 at 5-minute and 15-minute aggregation levels.
The auto correlations described by Figure 6-1 are typical of the volume series at other stations. These Auto Correlation Function (ACF) plots draw attention to an interesting feature that can be generalized to other locations. There are 288 5-minute intervals in a 24-hour period. The existence of non-stationarity due to periodic nature of 5-minute data is expected in the form of spikes at lags or periods at multiples of 288. At the 5-minute level, the long term daily (period = multiples of 288) and weekly (period= 2016) autocorrelations are not significant with simple differencing at lag 1 (refer Figure 6-1a1). Spikes in ACF can be noticed only at low lags. With a combined weekly and simple differencing (differencing at lag =1 \textit{and} lag = 288 \textit{or} 2016), a significant spike is noticed at low lags and at the lag of seasonal difference (288 or 2016), see Figure 6-1a4, Figure 6-1a5. Therefore, for 5-minute data, a simple differencing might be sufficient to yield an approximately stationary series. However, for the 15 minute aggregated data, it can be seen that there are significant spikes at daily (period = multiples of 96) and weekly lags (period =672), even for simple differencing of lag 1, see Figure 6-1b1. Seasonal or weekly differencing is warranted, at daily or weekly lags to achieve an exponentially declining ACF function for stationarity, resulting in Seasonal ARIMA models. Therefore, the need for seasonal differencing increases (to make the series stationary) as the aggregation level changes from 5 minutes, to 10 minutes to 15 minutes or further. At aggregation levels of 5 minutes, volume series can be simply differenced to achieve stationarity. At \geq 15-minutes aggregation levels, volume series need to be differenced seasonally –at either daily / weekly lag (lag=96/672 for 15 minute aggregation) to achieve stationarity.

Another instance is given by the ACF plots of simple differenced volumes at different aggregation levels at station 60, in Figure 6-2. The seasonal correlations are
non-significant for 5 minute volumes, vague for 10 minute volumes but significant and clear for 15 minute volumes and higher.

![ACF plots](image)

**Figure 6-2: ACF of simply differenced volumes at different aggregation levels at station 60**

Following the above discussion, it makes sense to analyze the cross correlations between upstream and downstream stations at 5-minute and 15-minute aggregation levels. This is not to look for long term correlations among the stations in terms of weekly / daily lags, but to see if the nature of the short term correlations among the stations changes with aggregation levels. Due to the fact that the stations are closely spaced, any traffic shocks / disturbances may dissipate when we aggregate at low levels (> 15 minutes).
Figure 6-3 shows the long term Cross-Correlation Functions (CCF) between volumes at station 35 and at station 34 at 5-minute and 15-minute aggregation levels. It can be seen from these figures that at 5-minute aggregation, a simple differencing operation is enough to control long term seasonal cross correlations (first figure from top in Figure 6-3a). For the 15-minute aggregated volumes, however, significant spikes remain at daily lags (lag = multiples of 96) after simple differencing (first row in Figure 6-3b). A second order differencing, with one simple and one at daily / weekly lag (lag = 96 / 672) is required to make the series stationary (fourth and fifth figures from top in Figure 6-3b).

![CCF for volume at 35 with 34 differenced at lag 1](image1)

![CCF for volume at 35 with 34 differenced at lag 288](image2)

![CCF for volume at 35 with 34 differenced at lag 2016](image3)

![CCF for volume at 35 with 34 differenced at lag 1 and 288](image4)

![CCF for volume at 35 with 34 differenced at lag 1 and 2016](image5)

**Figure 6-3 a. Cross-correlations between volumes on station 35 and station 34 at 5-minute aggregation**
Figure 6-3 bCross-correlations between volumes on station 35 and station 34 at 15-minute aggregation

Figure 6-3: Cross-correlations between volumes on Station 35 and volumes on Station 34 at 5-minute and 15-minute aggregation levels, up to lags corresponding to a week

The cross correlations between simple differenced volumes for adjacent stations at two different aggregation levels as shown in Figure 6-3, visually points to an interesting feature, that is consistent with the ACF function for individual stations – at 5-minute aggregation level, the cross correlation at low lags is significant (spikes closer to lag =0). For 15-minute aggregation, the cross correlation seems to be significant at daily lags (multiples of 96). This behavior is consistent across the stations in the study area. When simple and seasonal differencing is combined, the cross correlations at low lags and the period of seasonal difference are significant for 5 minute aggregated data, while the cross correlations at period of seasonal difference are more significant for 15-minute
aggregated data, with some marginally significant cross correlations present at low lags around zero. Therefore, as the aggregation level increases, the correlations from immediate neighborhood are expected to decline. Whenever cross correlations at both positive and negative lags are significant, feedback processes are involved and the effect of both upstream and downstream locations has to be considered simultaneously. In such cases, a univariate time-series methodology is inappropriate and a multivariate methodology is necessary. The cross correlations show that multivariate methods might be better at prediction of traffic variables at lower aggregation levels (say 5-minute aggregation) than higher (15-minute aggregation).

The Vector Autoregressive (VAR) models exploit the dynamic interactions among interrelated time series. For traffic data arising from numerous detectors on a stretch of freeway, VAR models can be applied to identify and quantify the effect of traffic at different time lags from other locations. VAR models can be used to describe and predict multiple spatially referenced time series as their formulation fits very well to the current study.

A VAR process of autoregressive order p is written as

\[ y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t \]

where \( y_t = (y_{1t}, \ldots, y_{kt})' \) \( t = 0,1,\ldots \) denotes a \( k \)-dimensional time series vector which are time series of observations from \( k \) loop detectors under consideration, \( \varepsilon_t \) is a vector of white noise process with \( (\varepsilon_{1t}, \varepsilon_{2t}, \varepsilon_{3t}, \ldots, \varepsilon_{kt})' \) and \( \varphi_1, \varphi_2, \ldots, \varphi_p \) represent parameter matrices that have to be estimated. SAS software was used to estimate VAR models.
For the current problem, different VAR models of appropriate order $p$ are fit for volumes from station 35 using other traffic variables from station 35 (speed, occupancy) and the surrounding stations based on the minimum Schwarz Bayesian Criterion (SBC). Figure 6-4 shows the cross correlations between volumes, occupancies and speeds at station 35 (pairwise). From the figures, it can be seen that volumes have a greater effect on future values of speeds and occupancies than vice versa (spikes more significant on the side of volumes than occupancies or speeds).

![CCF between Occupancy (response) and Volume (input)](image)

Figure 6-4a: Cross correlation between occupancy and volume at station 35
Figure 6-4b: Cross correlation between speed and volume at station 35

Figure 6-4c: Cross correlation between speed and occupancy at station 35

Figure 6-4: Cross correlations between Occupancy, Volume and Speeds at station 35
An unrestricted VAR model is fit first and then the insignificant parameters are constrained to zero.

The equation below shows the significant parameters in the VAR(p) model for all the three stations, with the insignificant coefficients constrained to zero. From the model, it can be seen that the significant parameters for volume at station 35 are both from station 34 (upstream to station 35) and station 36 (downstream to station 35). From the VAR model univariate equations can be written for each location as a function of past values of itself and other locations. The multivariate model for the differentiated volumes for station 35, 34 (upstream to 35) and 36 (downstream to 35) can now be written as

\[
\begin{bmatrix}
V_{35t}
\end{bmatrix} = \begin{bmatrix}
-0.61290 & 0.13184 & 0.14532
\end{bmatrix} + \begin{bmatrix}
V_{35t-1}
\end{bmatrix} + \begin{bmatrix}
V_{36t-1}
\end{bmatrix} + \begin{bmatrix}
V_{36t-2}
\end{bmatrix} + \begin{bmatrix}
\epsilon_{35t}
\end{bmatrix}
\]

where \( V_{35t} = V_{35t} - V_{35t-1}, V_{34t} = V_{34t} - V_{34t-1}, V_{36t} = V_{36t} - V_{36t-1} \)

\( V_{xt} \) : Volume at station \( x \) at time \( t \), \( \epsilon_{xt} \) : error from the model for station \( x \) at time \( t \)

The above formulation is equivalent to estimating three different univariate equations for volumes at stations 34, 35 and 36 and adding a condition that relates the error covariances across these variables. However, if univariate ARIMA models were fit, the cross dependency among the variables would have been ignored.

Similarly, a VAR model with volume, speed and occupancy at station 35 as the endogenous variables is estimated as

\[
\begin{bmatrix}
V_{35t}
\end{bmatrix} = \begin{bmatrix}
-0.43871 & 0.00000 & -0.55448
\end{bmatrix} + \begin{bmatrix}
V_{35t-1}
\end{bmatrix} + \begin{bmatrix}
V_{36t-1}
\end{bmatrix} + \begin{bmatrix}
V_{36t-2}
\end{bmatrix} + \begin{bmatrix}
\epsilon_{35t}
\end{bmatrix}
\]

where \( V_{xt} = V_{xt} - V_{xt-1} \) : simple differenced volume at station \( x \) at time \( t \)
\[ s_{xt} = S_{xt} - S_{xt-1} \]: simple differenced speed at station \( x \) at time \( t \)

\[ o_{xt} = O_{xt} - O_{xt-1} \]: simple differenced occupancy at station \( x \) at time \( t \)

\( V_{xt} \): Volume at station \( x \) at time \( t \)

\( S_{xt} \): Speed at station \( x \) at time \( t \)

\( O_{xt} \): Occupancy at station \( x \) at time \( t \)

\( \varepsilon^v_{xt} \): error from the model for volume for station \( x \) at time \( t \)

\( \varepsilon^s_{xt} \): error from the model for speed for station \( x \) at time \( t \)

\( \varepsilon^o_{xt} \): error from the model for occupancy for station \( x \) at time \( t \)

This single model provides three univariate equations of the three variables simultaneously and therefore, can be used to forecast speeds, occupancies and volumes, based on the past values of all other variables. In Eq-2, the signs and values of the parameters provide an insight into the behavior of the multivariate system. For instance, the change in volume at a time step \( t \) \((v_t)\) depends on the change in volume in the past 3 time steps \((t-1, t-2, t-3)\), the change in occupancy in the immediate past time step \((t-1)\), and the change in speed after a delay of one time period \((t-2, t-3)\). The univariate models for the change in volume, speed and occupancy can be written as

\[ v_{35t} = (-0.43871) * v_{35t-1} + (-0.55448) * o_{35t-1} + (-0.16683) * v_{35t-2} + 0.62599 * s_{35t-2} + (0.06155) * v_{35t-3} + 0.28439 * s_{35t-3} \ldots \text{Eq 3} \]

\[ s_{35t} = (-0.00991) * v_{35t-1} + (-0.04377) * o_{35t-1} + (-0.00993) * v_{35t-2} + (-0.07222) * o_{35t-2} + (-0.00683) * v_{35t-3} \]
\[ (++-0.06069) * o_{35t-3} \ldots \text{Eq 4} \]

\[ o_{35t} = 0.00513 * v_{35t-1} + (-0.34065) * s_{35t-1} + (-0.40898) * o_{35t-1} + 0.00555 * v_{35t-2} + (-0.15073) * v_{35t-2} \]
\[ (+0.09418) * s_{35t-3} + (-0.05984) * o_{35t-3} \ldots \text{Eq 5} \]

This shows that an increase in volume in one time step will be accompanied by a decrease in the next time period, while a decrease in volume in one time step will be accompanied by an increase in the next time step (due to the negative signs of \( v_{35t-1}, v_{35t-2} \) in the equation). An increase in occupancy at any time will lead in a decrease in volume,
while an increase in speed will lead to increase in volume. Also, it is interesting to note that the future speeds are sensitive to past volumes and occupancies. A similar reasoning can be applied to Eq(1), to measure the effects of sudden increases in volumes in station 34, on station 35 and station 36. This formulation enables the analyst to measure the effect of increase / decrease of the input variable on the future value of an output variable. In fact, such analysis can be formally carried out by Impulse Response Analysis (IRA) of VAR models, which traces the effect of shocks on the whole VAR system variables. When applied to spatial time series variables in VAR formulation, IRA can provide valuable insights into the effects of “shocks” or impulses in a variable on the future values of other variables (e.g., what is the effect of increase in volume at station 35 on volumes at station 34 and station 36). Such an analysis would provide greater management capability to traffic engineers and operators during incident management.

Similar models can be formulated for station 34 and station 36, based on coefficients that relate the response station to the input station. The model was then expanded to include two stations upstream (station 33, station 34) and two stations downstream (station 36, station 37) to station 35 and appropriate VAR models were fit for 5-minutes, 10-minutes and 15-minutes aggregated data. A plot of the cross correlations from the residuals for station 35, with its neighboring stations is shown in Figure 6-5(for 5-minute aggregated data) and Figure 6-6(for 15-minute aggregated data).
Figure 6-5: Cross correlation among the residuals of various VAR models at 5-minute aggregation
Figure 6-6: Cross correlation among the residuals of various VAR models at 15-minute aggregation
From Figure 6-5 and Figure 6-6 some observations regarding the performance of the models can be made. From Figure 6-5, at 5-minute aggregation, the VAR models with volume, speed and occupancy from station 35 as endogenous variables perform reasonably in capturing the correlations between the volume and speed (cross correlations with occupancy not shown in the figures) at very low lags (sub-figure in first row, left column in Figure 6-5). When VAR models are fit with volumes from immediate neighbors (station 34, station 35, station 36) as endogenous variables, there are no significant spikes in cross correlation between volumes from station 35 with station 34 or with station 36 at short lags (refer to second and third rows in Figure 6-4). The addition of two upstream and downstream volumes (station 33, 34; station 36, 37) also suppresses any residual cross correlation between station 35 with its upstream (station 33, 34) and downstream (station 36, 37) at short (lags <=30) as well as long lags, (refer to last four rows in Figure 6-5). The spikes at daily weekly lags are not very pronounced (from the right side figures in Figure 6-5. From Figure 6-6, the VAR models for 15-minute aggregated data capture short term cross correlations, but the spikes at daily lags (multiples of 96) are more pronounced on the figures on the right side. This can be explained from the fact that at lower time scales (aggregation at 5-minutes say), the variation is due to local conditions, and can be explained better with the surrounding spatio-temporal data. As the aggregation time scale increases (15-minutes, say), the data is de-noised, and deviates less from the local historical values. Inter spatial dependencies are captured into the cross correlation at lag 0 (the big spike in the ccf plot in the middle) at higher time scales of aggregation. Therefore, it can be postulated that as aggregation of data happens at larger time scales, univariate time series models might provide a better representation and prediction of the process.
The advantage with a VAR model is also that one model can estimate the coefficients (parameter matrices) in one run. However, as the number of series increases, the number of parameters increases exponentially. To overcome this, a structure for the covariance matrix can be assumed that translates into constraining some coefficients in the parameter matrices. This is a non-trivial issue, as the spatio-temporal characteristics of traffic are likely to change throughout the day, during rush hours or incident conditions, etc. However, occupancy has traditionally been used as a proxy indicator for traffic congestion, and can be used as another variable in the multivariate formulation. A complete multivariate specification can therefore include all variables from the station (volumes, speeds, occupancies) as well as its neighbors. This requires future investigation.

Table 6-1 shows the comparison of the univariate ARIMA models following Box – Jenkins methodology with the multivariate VAR models based on the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) for station 35. A simple ARIMA (p,d,q) and a Seasonal ARIMA (p,d,q) \times (P,D,Q) s are chosen as the candidate univariate models. Details of these models can be inferred from Ahmed and Cook (1979) and Williams, et al. (1998). Table 6-2 shows the comparison of univariate and multivariate time series prediction models for station 8 and station 60 in the eastbound direction. The training data set is the set of days from March 1st, 2003 to March 24th, 2003. The test data set is the week from March 25th, 2003 to March 31st, 2003. RMSE and MAPE are defined as

\[
RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N}} \quad \text{MAPE} = \frac{1}{N} \sum |y_i - \hat{y}_i|
\]

where \(y_i = \) observed value of the response \(\hat{y}_i = \) forecasts for \(y_i\)
Table 6-1: Measures of Performance for ARIMA, VAR Models for Volumes at Station 35

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Model</th>
<th>Endogenous Variables</th>
<th>Difference</th>
<th>Training RMSE</th>
<th>Training MAPE</th>
<th>Test RMSE</th>
<th>Test MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume</td>
<td>ARIMA(0,1,2)</td>
<td>None</td>
<td>1</td>
<td>30.21</td>
<td>9.73</td>
<td>30.11</td>
<td>9.93</td>
</tr>
<tr>
<td>Volume</td>
<td>SARIMA(0,1,2)X(0,1,1) s=288</td>
<td>None</td>
<td>1,288</td>
<td>31.42</td>
<td>10.37</td>
<td>30.86</td>
<td>10.42</td>
</tr>
<tr>
<td>Volume</td>
<td>SARIMA(0,1,2)X(0,1,1) s=2016</td>
<td>None</td>
<td>1,2016</td>
<td>33.78</td>
<td>11.65</td>
<td>36.23</td>
<td>12.86</td>
</tr>
<tr>
<td>Volume</td>
<td>VAR(3)</td>
<td>Speed at 35, Occupancy at 35</td>
<td>1</td>
<td>28.15</td>
<td>9.73</td>
<td>28.85</td>
<td>9.9</td>
</tr>
<tr>
<td>Volume</td>
<td>VAR(3)</td>
<td>Speed at 35, Occupancy at 35</td>
<td>1,288</td>
<td>33.26</td>
<td>12.05</td>
<td>37.11</td>
<td>12.72</td>
</tr>
<tr>
<td>Volume</td>
<td>VAR(4)</td>
<td>Volume at 34, Volume at 36</td>
<td>1</td>
<td>27.72</td>
<td>9.57</td>
<td>28.39</td>
<td>9.78</td>
</tr>
<tr>
<td>Volume</td>
<td>VAR(6)</td>
<td>Volumes at Stations 33,34,36,37</td>
<td>1</td>
<td>26.66</td>
<td>9.4</td>
<td>27.9</td>
<td>9.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Model</th>
<th>Endogenous Variables</th>
<th>Difference</th>
<th>Training RMSE</th>
<th>Training MAPE</th>
<th>Test RMSE</th>
<th>Test MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume</td>
<td>ARIMA(0,1,6)</td>
<td>None</td>
<td>1</td>
<td>54.58</td>
<td>10.7</td>
<td>57.93</td>
<td>10.35</td>
</tr>
<tr>
<td>Volume</td>
<td>SARIMA(1,1,2)X(0,1,1) s=144</td>
<td>None</td>
<td>1,144</td>
<td>53.29</td>
<td>10.18</td>
<td>54.72</td>
<td>9.94</td>
</tr>
<tr>
<td>Volume</td>
<td>SARIMA(1,1,1)X(0,1,1) s=1008</td>
<td>None</td>
<td>1,1008</td>
<td>57.15</td>
<td>9.6</td>
<td>61.21</td>
<td>12.11</td>
</tr>
<tr>
<td>Volume</td>
<td>VAR(3)</td>
<td>Speed at 35, Occupancy at 35</td>
<td>1</td>
<td>52.84</td>
<td>9.01</td>
<td>56.08</td>
<td>9.36</td>
</tr>
<tr>
<td>Volume</td>
<td>VAR(3)</td>
<td>Speed at 35, Occupancy at 35</td>
<td>1,144</td>
<td>59.49</td>
<td>10.35</td>
<td>70.2</td>
<td>11.45</td>
</tr>
<tr>
<td>Volume</td>
<td>VAR(6)</td>
<td>Volume at 34, Volume at 36</td>
<td>1</td>
<td>51.46</td>
<td>8.77</td>
<td>50.87</td>
<td>8.22</td>
</tr>
<tr>
<td>Volume</td>
<td>VAR(6)</td>
<td>Volumes at Stations 33,34,36,37</td>
<td>1</td>
<td>50.36</td>
<td>8.34</td>
<td>51.38</td>
<td>8.53</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Model</th>
<th>Endogenous Variables</th>
<th>Difference</th>
<th>Training RMSE</th>
<th>Training MAPE</th>
<th>Test RMSE</th>
<th>Test MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume</td>
<td>ARIMA(2,1,2)</td>
<td>None</td>
<td>1</td>
<td>81.77</td>
<td>9.21</td>
<td>89.06</td>
<td>9.62</td>
</tr>
<tr>
<td>Volume</td>
<td>SARIMA(1,1,2)X(0,1,1) s=96</td>
<td>None</td>
<td>1,96</td>
<td>80.33</td>
<td>9.32</td>
<td>78.49</td>
<td>8.96</td>
</tr>
<tr>
<td>Volume</td>
<td>SARIMA(1,1,1)X(0,1,1) s=672</td>
<td>None</td>
<td>1,672</td>
<td>82.13</td>
<td>9.18</td>
<td>88.45</td>
<td>10.22</td>
</tr>
<tr>
<td>Volume</td>
<td>VAR(3)</td>
<td>Speed at 35, Occupancy at 35</td>
<td>1</td>
<td>82.18</td>
<td>9.38</td>
<td>84.13</td>
<td>9.53</td>
</tr>
<tr>
<td>Volume</td>
<td>VAR(3)</td>
<td>Speed at 35, Occupancy at 35</td>
<td>1,96</td>
<td>90.1</td>
<td>10.55</td>
<td>92.11</td>
<td>10.11</td>
</tr>
<tr>
<td>Volume</td>
<td>VAR(6)</td>
<td>Volume at 34, Volume at 36</td>
<td>1</td>
<td>83.22</td>
<td>9.36</td>
<td>85.22</td>
<td>9.66</td>
</tr>
<tr>
<td>Volume</td>
<td>VAR(6)</td>
<td>Volumes at Stations 33,34,36,37</td>
<td>1</td>
<td>78.19</td>
<td>8.77</td>
<td>81.06</td>
<td>9.36</td>
</tr>
</tbody>
</table>

From Table 6-1, it can be seen that VAR models have a lower RMSE and MAPE than the univariate ARIMA models for the training as well as test data sets for stations 34, 35 and...
Specifically, at the aggregation level of 5-minutes, models with a simple difference (difference = 1) are better than models with combined simple seasonal differences (difference = 1, 288 / 2016). VAR models perform better than ARIMA models, in terms of RMSE and MAPE on training and test data sets. The VAR model with 2 neighbors performed the best. At 10-minute aggregation level, the SARIMA model with simple and daily difference (difference lag = 1, 96) outperformed the simple ARIMA model. VAR models with the same differencing (1, 96) suffered in performance. However, VAR models with simple difference with one and two neighbors were still the best models among all candidate models. At 15-minute aggregation, the SARIMA model with simple and daily seasonal differencing (at lags =1, 96) was the best model. While some VAR models were arguably slightly better for the training set, they didn’t perform as well over the test data. Thus, SARIMA model was more robust at 15-minute level of aggregation.

The RMSE increased between the aggregation levels (from 5 to 10 to 15 minutes) for all models, because the scale of measure of volume increased. In effect, it is expected that 10 minute volumes will roughly be twice that of 5-minute volumes, which doubles the error. This is the reason for increase in RMSE for the same class of models (ARIMA, SARIMA and VAR) across aggregation levels. The VAR models consistently had lower RMSE when compared to the univariate models for 5-minute and 10-minute aggregation levels. The performance becomes more competitive at 15-minute aggregation levels.

Table 6-2 shows similar performances of the multivariate models for two other locations – station 8, and station 60 in the eastbound direction on I-4. It can be seen from these results also that multivariate time series prediction models perform better than univariate ARIMA models at low time scales of aggregation. As the time scale of aggregation becomes larger (say 5 minutes to 15 minutes), Seasonal ARIMA models
provide greater accuracy consistently, in case of traffic flows. However, at higher time scales of aggregation (15 minutes), VAR models might not be as robust as SARIMA models. One advantage of VAR estimation is that multiple dependent variables (speeds, volumes, occupancies) can be estimated simultaneously. Moreover, a VAR model provides insight into the impacts of change in one variable on all other variables. Thus, VAR models provide framework for answering how a drop in volume in the current period could impact speeds (or neighboring volumes) in the future periods, through Impulse Response Functions (Luktepohl 1993). This aspect is the subject of future research.

A complete specification of a multivariate traffic prediction problem should ideally incorporate multivariable relationships that exist among different traffic variables in different locations. However, this is a non-trivial problem (and the model can suffer from over parameterization) and at large enough time scales, the correlations among different variables tend to be insignificant. The results in this chapter have shown that while multivariate methods might be suitable for prediction and could outperform univariate methods, their accuracy can be questionable at higher time scales (>=15 minute intervals). Then, the multivariate traffic prediction would reduce to univariate models, if all the insignificant variables are constrained to zeros. On the other hand, multivariable methods at low aggregations can provide system wide forecasts with relatively simple formulations. This could be effectively used to post predictive traffic information on websites, in-vehicle GPS systems, and DMS signs.
Table 6-2: Measures of Performance for ARIMA, VAR Models for Volumes at Stations 8 and 60 at aggregation levels of 5 / 10 / 15 minutes.

<table>
<thead>
<tr>
<th>Aggregation - 5 minutes</th>
<th>Station</th>
<th>Model</th>
<th>RMSE Train</th>
<th>RMSE Test</th>
<th>MAPE Train</th>
<th>MAPE Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
<td>ARIMA</td>
<td>24.9</td>
<td>24.37</td>
<td>9.71</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>SARIMA</td>
<td>25.5</td>
<td>24.31</td>
<td>11.87</td>
<td>11.38</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>VAR0</td>
<td>23.53</td>
<td>22.27</td>
<td>9.71</td>
<td>9.41</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>VAR1</td>
<td>22.72</td>
<td>22.69</td>
<td>9.31</td>
<td>9.39</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>VAR2</td>
<td>21.5</td>
<td>22.33</td>
<td>9.22</td>
<td>9.02</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>ARIMA</td>
<td>26.34</td>
<td>26.21</td>
<td>10.38</td>
<td>10.17</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>SARIMA</td>
<td>25.6</td>
<td>24.02</td>
<td>10.67</td>
<td>10.26</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>VAR0</td>
<td>23.21</td>
<td>23.18</td>
<td>10.39</td>
<td>10.33</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>VAR1</td>
<td>22.02</td>
<td>21.15</td>
<td>10.24</td>
<td>9.86</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>VAR2</td>
<td>20.98</td>
<td>21</td>
<td>9.69</td>
<td>9.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Aggregation - 10 minutes</th>
<th>Station</th>
<th>Model</th>
<th>RMSE Train</th>
<th>RMSE Test</th>
<th>MAPE Train</th>
<th>MAPE Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
<td>ARIMA</td>
<td>39.89</td>
<td>39.21</td>
<td>7.83</td>
<td>8.23</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>SARIMA</td>
<td>38.56</td>
<td>38.04</td>
<td>7.67</td>
<td>7.86</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>VAR0</td>
<td>40.05</td>
<td>40.86</td>
<td>8.27</td>
<td>8.16</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>VAR1</td>
<td>36.82</td>
<td>37.33</td>
<td>7.71</td>
<td>7.26</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>VAR2</td>
<td>36.16</td>
<td>36.47</td>
<td>7.73</td>
<td>7.22</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>ARIMA</td>
<td>38.83</td>
<td>38.98</td>
<td>8.75</td>
<td>8.66</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>SARIMA</td>
<td>37.08</td>
<td>36.87</td>
<td>8.63</td>
<td>9.12</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>VAR0</td>
<td>39.01</td>
<td>38.45</td>
<td>8.93</td>
<td>9.87</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>VAR1</td>
<td>36.02</td>
<td>35.21</td>
<td>7.87</td>
<td>7.53</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>VAR2</td>
<td>35.44</td>
<td>35.03</td>
<td>7.13</td>
<td>7.87</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Aggregation - 15 minutes</th>
<th>Station</th>
<th>Model</th>
<th>RMSE Train</th>
<th>RMSE Test</th>
<th>MAPE Train</th>
<th>MAPE Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
<td>ARIMA</td>
<td>59.92</td>
<td>61.86</td>
<td>8.29</td>
<td>8.11</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>SARIMA</td>
<td>51.81</td>
<td>57.5</td>
<td>6.97</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>VAR0</td>
<td>56.65</td>
<td>55.87</td>
<td>7.55</td>
<td>7.32</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>VAR1</td>
<td>55.73</td>
<td>57.23</td>
<td>7.43</td>
<td>7.87</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>VAR2</td>
<td>54.81</td>
<td>56.86</td>
<td>7.36</td>
<td>8.03</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>ARIMA</td>
<td>56.08</td>
<td>56.72</td>
<td>8.44</td>
<td>8.41</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>SARIMA</td>
<td>51.71</td>
<td>51.35</td>
<td>7.98</td>
<td>8.03</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>VAR0</td>
<td>56.42</td>
<td>60.86</td>
<td>8.52</td>
<td>9.15</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>VAR1</td>
<td>53.54</td>
<td>55.23</td>
<td>8.97</td>
<td>8.84</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>VAR2</td>
<td>52.12</td>
<td>57.26</td>
<td>8.71</td>
<td>8.45</td>
</tr>
</tbody>
</table>

VAR0: VAR model with speed, occupancy and volume at the same station as endogenous variables
VAR1: VAR model with volumes from 1 upstream and downstream neighbor as endogenous variables
VAR2: VAR model with volumes from 2 upstream and downstream neighbors as endogenous variables
6.5 Conclusions

In this chapter, multivariate time series models for prediction were estimated at three locations at three aggregation levels – 5 minutes, 10 minute and 15 minutes. For the purpose of demonstration, volumes from station 35 on the eastbound direction of I-4, Orlando were used in correlation and cross correlation analysis at three different aggregation time scales. The graphical analysis showed that short term cross correlations tend to be stronger when the volumes are aggregated at 5 minute intervals. The daily and weekly (day of the week) autocorrelations become stronger when aggregation takes place at 10 or 15 minute intervals. At 10 minute intervals, the weekly correlations and cross correlations were not as profound as for 15 minute aggregation, and not as insignificant as for 5 minute aggregation. This could have a significant effect in determining the effectiveness of seasonally differenced models at respective aggregation levels.

VAR analysis was performed to show that VAR models generally outperform univariate ARIMA and Seasonal ARIMA (SARIMA) models for prediction at data aggregation levels of 5 minutes and 10 minutes. It has been shown that VAR models have lower errors than ARIMA models on both the training and test data sets. The residuals from the VAR models do not have significant cross correlations at 5 minute levels – this implies that no further information can be extracted from the neighboring upstream and downstream stations for prediction. At 15 minute aggregation levels VAR models did not outperform SARIMA models, and the residuals showed strong correlations at weekly lags. VAR models may not be justified for prediction at aggregation levels higher than 15 minutes, as local disturbances among locations in close proximity tend to disappear within the same time interval (if the time interval is wide enough, for example 15 minutes or greater). In such cases, a larger spatial neighborhood
can be considered for prediction. However, due to wide spatial and temporal extents, feedback would not be significant and transfer function models for prediction could suffice (that consider upstream effects only). Further investigation of VAR models is required at higher time scales (30 minutes, 60 minutes).

The chapter has demonstrated cross correlation analysis and showed that multivariate spatial time series on a freeway section can be forecasted using VAR models at data resolutions < 15 minutes, without loss of accuracy. The research introduced in this chapter can be further refined as described earlier to result in better prediction models for freeway traffic data.
CHAPTER 7. CONCLUSIONS AND FUTURE WORK

The research in this dissertation is to improve freeway traffic prediction by the investigation of spatial and temporal relationships among the data through multivariate cross correlations and multivariate time series models. Traffic prediction is an important component of the Advanced Traffic Information System (ATIS). As the demand increases on the roads, the need for real-time and predictive traffic information becomes critical. Freeway traffic exhibits strong autocorrelation in the spatial and temporal domains, which should be utilized for a more efficient traffic prediction.

Throughout the course of this dissertation, a case has been made about the utilization of spatial nature of traffic, particularly with respect to freeway traffic data and its prediction. While most of the traditional statistical approaches have incorporated temporal covariances, inter-location correlations have not been extensively investigated. In spite of the complexity associated with defining model structures, spatial information adds to our knowledge of the process and provides us with more confidence in our inferences.

Univariate time series models were built for specific locations for each traffic variable, and along-side them multivariate time series models for spatial time series were also built. By incorporating the cross correlations between every pair of locations in the study region, a forecasting model was cast for each variable in question. It was proved that multivariate approach to spatial data would provide for a richer understanding of the traffic process from a statistical as well as physical point of view.
7.1 Summary of Research

Cross covariances and cross correlations were used to characterize the relationships between the variables that were collected at different locations. Specifically, it was shown that within the extent of 3 miles (about 5 stations), there were significant cross correlations between stations for each variable - speed, volume and occupancy. At any location, there were also cross correlations between speeds, volumes and occupancies. The nature of these cross correlations is such that any variable influences future values of other variables and is influenced by the past values of other variables. Feedback is significant, especially for immediate upstream / downstream stations.

The formulation of Vector Autoregressive (VAR) model lends itself well to the analysis of system of variables that influence each other. The VAR model has been extensively used in the analysis of economic time series, and provides a sound framework for analyzing inter-variable effects. They are extensions of univariate time series methods into multivariate domain.

VAR models were applied to speeds, volumes and occupancies and were compared with univariate time series counterparts (specifically for speeds and volumes). In the case of speeds and volume predictions at 5-minute intervals, the VAR models outperformed the univariate methods, at various locations (i.e., loop detector stations). The complexity of VAR models was an issue, but it must be understood that VAR models provide predictions for multiple variables simultaneously. For a VAR with $k$ variables, their complexity must be compared with $k$ univariate models. While VAR models were not simpler models than the univariate counterparts, they were not too complex as was evidenced by the information criteria (SBC). In addition, they provided with the effects of upstream speeds and volumes on downstream locations and vice versa.
Such an analysis can be very critical for Advanced Traffic Management System, when an unexpected incident can introduce random traffic shocks on the segments.

The performance of the models was also tested at different aggregation levels – 5 minutes, 10 minutes and 15 minutes and the results were reported for these levels. As the data aggregation was performed at higher time scales, the performance of VAR model prediction suffered when compared to univariate models. At aggregation levels of 15 minutes (and greater), a seasonal ARIMA model performed better for speeds and volumes. This is because when data are aggregated at higher time scales, the variation between the locations gets dampened. However, for aggregation levels < 15 minutes, a VAR model for close spatial neighbors works better than ARIMA models.

With the ubiquity of various ITS technologies like Dynamic Message Signs (DMS), Internet enabled phones, and Personal Navigation Devices (PNDs), there is an increasing demand for more accurate and reliable information. Real-time information is most accurate, but there is an increasing interest and competition among providers to disseminate reliable predictive traffic information to users. The economical and ecological environment also calls for a better management of the transportation system and transportation choices, in terms of trip time and trip mode. In such a climate a more responsive traffic prediction system is needed.

Advanced Traffic Management Systems (ATMS) rely on traffic redistribution and redirection throughout the network. With the current travel time algorithms that do not use spatial components, it is difficult to predict the effects of diversion or rerouting vehicles at a specific location upstream or downstream of the re-routing point. VAR models provide the framework to answer questions on such effects, as they incorporate
multivariate relationships among data that are not provided by univariate prediction models.

The DMS most frequently provide real-time and predictive travel time information, based on algorithms that model travel times from real-time traffic and forecast future travel times based on forecasted / predicted traffic information. Real-time information can be reliable for short segments, but as the segment length increases, the difference between real-time and predictive information becomes fuzzy. The DMS is one ATIS technology that can utilize algorithms to utilize improved traffic prediction algorithms. The credibility of the messages on DMS is directly related to the quality of traffic forecasting algorithms used. An improved traffic prediction algorithm can improve travel time estimates based on predictive traffic information, and therefore improve the credibility of the system.

Some traffic devices (Blackberry, etc.) have the capability to receive pre-trip or en-route real-time / predictive traffic information. This information can be used to schedule trips by commuters to optimize their trips such that they avoid the congested times and routes. When the information is to be used for pre-trip scheduling or en-route rescheduling, the predictive traffic information needs to be reliable. For such an application, accurate and reliable predicted traffic information that takes spatial and temporal dynamics into account forms the basis for its utility.

A lot of traffic information is posted on the web sites that provide real time and predictive traffic information (www.georgianavigator.com ...etc). With the advent of traffic data warehousing, data mining, and the accessibility of internet, commuters refer to the internet for predictive traffic information to schedule their trips. For such traffic
information websites, the predicted travel times and predicted speeds are the basis of the applications.

With the IntelliDrive program, individual vehicles will provide the RTMC (Regional Traffic Management Center) with traffic and vehicle data on a continuous basis. This would provide instantaneous travel time and traffic information, as well as actual “in the field” travel times. While the data used in this dissertation is from the loop detector data, the basic framework can be applied to any form of data that is collected at a temporal and spatial scale. However, it would need to be pre-processed and aggregated suitably for the methods to be applicable.

7.2 Recommendations for Future Work

This dissertation adds to the body of knowledge on freeway traffic forecasting problem by proposing new multivariate techniques for prediction, at different time scales. The use of multivariate models will enhance the understanding of the traffic process. It has been demonstrated in the previous sections that using information from across space (station data) and across time can provide users with a better estimate of the future response within reasonable time scales.

It is expected that these procedures will shed more light on the understanding of the traffic phenomena. Moreover, the existence of multiple states in traffic can be modeled by assuming heteroscedastic models. These would mean that the variances and covariances change with time and therefore Multivariate GARCH models need to be assessed. These models require further investigation. Further, spatial structures based on the locations of loop detector stations need to be investigated. When the time scale of aggregation gets higher, locations farther out can become significant. The presence of
numerous ramps can dampen these effects. As such, these effects need to be investigated. However, the multivariate methods introduced in this dissertation are expected to provide a strong framework for all such future analyses.
APPENDIX A: SPEEDS SURROUNDING STATION 8 AND CORRESPONDING CCF AND ACF
APPENDIX B: VOLUMES SURROUNDING STATION 35 AND CORRESPONDING ACF AND CCF
APPENDIX C: RESIDUALS FROM ARIMA MODELS FOR STATIONS SURROUNDING STATION 60 AND THEIR CORRESPONDING CCF
APPENDIX D: COMPARISON OF VAR MODEL WITH ARIMA MODEL
# ARIMA Model for stations 5-11

<table>
<thead>
<tr>
<th>Station</th>
<th>Variable</th>
<th>Model</th>
<th>RMSE</th>
<th>MAPE</th>
<th>RMSE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Speed</td>
<td>ARIMA(0,1,q)</td>
<td>3.27</td>
<td>3.96</td>
<td>3.63</td>
<td>3.92</td>
</tr>
<tr>
<td>6</td>
<td>Speed</td>
<td>ARIMA(0,1,q)</td>
<td>2.54</td>
<td>3.29</td>
<td>2.97</td>
<td>3.54</td>
</tr>
<tr>
<td>7</td>
<td>Speed</td>
<td>ARIMA(0,1,q)</td>
<td>2.58</td>
<td>2.97</td>
<td>2.6</td>
<td>2.77</td>
</tr>
<tr>
<td>8</td>
<td>Speed</td>
<td>ARIMA(0,1,q)</td>
<td>2.57</td>
<td>2.98</td>
<td>2.87</td>
<td>3.03</td>
</tr>
<tr>
<td>9</td>
<td>Speed</td>
<td>ARIMA(0,1,q)</td>
<td>2.74</td>
<td>3.15</td>
<td>2.8</td>
<td>3.19</td>
</tr>
<tr>
<td>10</td>
<td>Speed</td>
<td>ARIMA(0,1,q)</td>
<td>2.98</td>
<td>3.99</td>
<td>3.73</td>
<td>4.33</td>
</tr>
<tr>
<td>11</td>
<td>Speed</td>
<td>ARIMA(0,1,q)</td>
<td>2.53</td>
<td>2.81</td>
<td>2.64</td>
<td>3.05</td>
</tr>
</tbody>
</table>
Comparison of VAR model with ARIMA Model
VAR Model for stations 5-11

<table>
<thead>
<tr>
<th>Station</th>
<th>Variable</th>
<th>VAR Model</th>
<th>Training RMSE</th>
<th>Training MAPE</th>
<th>Test RMSE</th>
<th>Test MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Speed</td>
<td>2 down</td>
<td>1.86</td>
<td>1.76</td>
<td>1.92</td>
<td>2.06</td>
</tr>
<tr>
<td>8</td>
<td>Speed</td>
<td>1 up 1 down</td>
<td>1.54</td>
<td>1.97</td>
<td>1.85</td>
<td>2.02</td>
</tr>
<tr>
<td>9</td>
<td>Speed</td>
<td>2 up</td>
<td>1.72</td>
<td>2.09</td>
<td>1.99</td>
<td>2.37</td>
</tr>
<tr>
<td>6</td>
<td>Speed</td>
<td>4 down</td>
<td>1.52</td>
<td>1.71</td>
<td>1.63</td>
<td>1.99</td>
</tr>
<tr>
<td>7</td>
<td>Speed</td>
<td>1 up 3 down</td>
<td>1.53</td>
<td>1.73</td>
<td>1.87</td>
<td>2.01</td>
</tr>
<tr>
<td>8</td>
<td>Speed</td>
<td>2 up 2 down</td>
<td>1.52</td>
<td>1.97</td>
<td>1.84</td>
<td>2.04</td>
</tr>
<tr>
<td>9</td>
<td>Speed</td>
<td>3 up 1 down</td>
<td>1.64</td>
<td>2.01</td>
<td>1.73</td>
<td>1.96</td>
</tr>
<tr>
<td>10</td>
<td>Speed</td>
<td>4 up</td>
<td>1.84</td>
<td>2.16</td>
<td>1.97</td>
<td>2.45</td>
</tr>
<tr>
<td>5</td>
<td>Speed</td>
<td>6 down</td>
<td>2.16</td>
<td>2.59</td>
<td>2.12</td>
<td>2.43</td>
</tr>
<tr>
<td>6</td>
<td>Speed</td>
<td>1 up 5 down</td>
<td>1.5</td>
<td>1.7</td>
<td>1.93</td>
<td>2.34</td>
</tr>
<tr>
<td>7</td>
<td>Speed</td>
<td>2 up 4 down</td>
<td>1.52</td>
<td>1.73</td>
<td>2.15</td>
<td>2.53</td>
</tr>
<tr>
<td>8</td>
<td>Speed</td>
<td>3 up 3 down</td>
<td>1.52</td>
<td>1.94</td>
<td>2.07</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>Speed</td>
<td>4 up 2 down</td>
<td>1.63</td>
<td>1.96</td>
<td>1.75</td>
<td>1.95</td>
</tr>
<tr>
<td>10</td>
<td>Speed</td>
<td>5 up 1 down</td>
<td>1.83</td>
<td>2.15</td>
<td>1.94</td>
<td>2.21</td>
</tr>
<tr>
<td>11</td>
<td>Speed</td>
<td>6 up</td>
<td>2.51</td>
<td>3.18</td>
<td>2.5</td>
<td>2.98</td>
</tr>
</tbody>
</table>
REFERENCES

2. “2007 Annual Urban Mobility Report”, Published by Texas Transportation Institute.


