FREEWAY SHORT-TERM TRAFFIC FLOW FORECASTING BY

CONSIDERING TRAFFIC VOLATILITY DYNAMICS AND MISSING DATA

SITUATIONS

A Thesis

by

YANRU ZHANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

August 2011

Major Subject: Civil Engineering

Freeway Short-term Traffic Flow Forecasting by Considering Traffic Volatility

Dynamics and Missing Data Situations

FREEWAY SHORT-TERM TRAFFIC FLOW FORECASTING BY

CONSIDERING TRAFFIC VOLATILITY DYNAMICS AND MISSING DATA

SITUATIONS

A Thesis

by

YANRU ZHANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,	Yunlong Zhang
Committee Members,	Dominique Lord
	Michael Sherman
Head of Department,	John Niedzwecki

August 2011

Major Subject: Civil Engineering

ABSTRACT

Freeway Short-term Traffic Flow Forecasting by Considering Traffic Volatility

Dynamics and Missing Data Situations. (August 2011)

Yanru Zhang, B.S., Beijing Jiaotong University

Chair of Advisory Committee: Dr. Yunlong Zhang

Short-term traffic flow forecasting is a critical function in advanced traffic

management systems (ATMS) and advanced traveler information systems (ATIS).

Accurate forecasting results are useful to indicate future traffic conditions and assist

traffic managers in seeking solutions to congestion problems on urban freeways and

surface streets. There is new research interest in short-term traffic flow forecasting due

to recent developments in ITS technologies. Previous research involves technologies in

multiple areas, and a significant number of forecasting methods exist in literature.

However, forecasting reliability is not properly addressed in existing studies. Most

forecasting methods only focus on the expected value of traffic flow, assuming constant

variance when perform forecasting. This method does not consider the volatility nature

of traffic flow data. This paper demonstrated that the variance part of traffic flow data is

not constant, and dependency exists. A volatility model studies the dependency among

the variance part of traffic flow data and provides a prediction range to indicate the

reliability of traffic flow forecasting. We proposed an ARIMA-GARCH (Autoregressive

Integrated Moving Average- AutoRegressive Conditional Heteroskedasticity) model to

study the volatile nature of traffic flow data. Another problem of existing studies is that most methods have limited forecasting abilities when there is missing data in historical or current traffic flow data. We developed a General Regression Neural Network(GRNN) based multivariate forecasting method to deal with this issue. This method uses upstream information to predict traffic flow at the studied site. The study results indicate that the ARIMA-GARCH model outperforms other methods in non-missing data situations, while the GRNN model performs better in missing data situations.

ACKNOWLEDGEMENTS

NOMENCLATURE

ARIMA	Autoregressive Integrated Moving Average
ATIS	Advanced Traveler Information Systems
ATMS	Advanced Traffic Management Systems
GARCH	AutoRegressive Conditional Heteroskedasticity
GRNN	General Regression Neural Network
HA	Historical Average
MAPE	Mean Absolute Percentage Error
RMSE	Root Mean Square Error
VAR	Vector AutoRegression

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION: THE IMPORTANCE OF RESEARCH

Traffic flow is the study of interactions among vehicles, drivers, and infrastructures. The major objective of traffic flow study is to understand and develop an optimal road network that can efficiently move traffic and ease traffic congestion. One major area in traffic flow study is the ability to forecast traffic flow in the next few minutes: in other words, short-term traffic flow forecasting. This section introduces the importance of this research, traffic stream properties and other critical issues related with short-term traffic flow forecasting.

Short-term traffic flow forecasting is a critical function in advanced traffic management systems (ATMS) and advanced traveler information systems (ATIS). Accurate forecasting results can indicate future traffic conditions, which support the development of proactive traffic control strategies in ATMS; provide real-time route guidance in ATIS; and evaluate proactive traffic control and real-time route guidance strategies, as well. Because traffic flow forecasting can assist in seeking solutions to traffic congestion on urban freeways and surface streets, there is new research interest in short-term traffic flow forecasting due to recent developments in ITS technologies.

This thesis follows the style of *Journal of Transportation Engineering*.

Vehicular traffic, as a stream or a continuum fluid, has several parameters associated with it: flow, density, and speed. These parameters provide information regarding the nature of traffic flow and are indicators that detect variations in traffic flow. Because a traffic stream is not uniform but varies over time and space, measurement of traffic flow is in fact the sampling of random variables. The forecasting result of traffic flow is not an absolute value, but estimated values based on experimental data. This research will use some statistical methods to analyze the traffic flow patterns and fit appropriate models based on the study of the underlining traffic flow patterns.

1.1 Traffic Stream Properties

Traffic flow (rate), speed, and density are three basic parameters that describe traffic conditions. The values of these parameters are crucial elements in evaluating the near future traffic conditions; thus, the predicted values assist traffic system operators and road users to modify their strategies in using the roadway system efficiently. One should have a brief knowledge of traffic flow parameters before study traffic flow forecasting methods. The following is a brief introduction of three fundamental traffic flow parameters: flow, speed, and density.

1.1.1 Flow

Typically, there are two ways of detecting the number of vehicles passing a certain point of the roadway: volume and flow rate. The Highway Capacity Manual 2000(HCM 2000) defines traffic volume as "the total number of vehicles that pass over a given point or section of a lane or roadway during a given time interval; volumes can be

expressed in terms of annual, daily, hourly, or sub-hourly periods." On the other hand, the traffic flow rate is defined as "the equivalent hourly rate at which vehicles pass over a given point or segment of a lane or roadway during a given time interval of less than 1 h, usually 15 min". Traffic volume reflects the actual number of vehicles been observed along a roadway during a certain time period. The time interval of the volume data can be larger than one hour. Traffic flow rate, different from traffic volume, is collected for intervals of less than one hour—usually fifteen minutes, and is expressed as vehicles per hour. In other words, traffic flow rate is not the actual number of vehicles observed on the roadway for an hour but "an equivalent hourly rate." Normally, volume and flow reflect traffic demand—the number of vehicles or drivers who desire to use a given roadway facility in a specific time interval. However, in near capacity situations, flow will be constrained by roadway capacity. Volumes will reflect capacity in this kind of situation.

Traffic volume varies in both time and space. Traffic volume obtained at different time intervals can be different. It can vary month-to-month, day-to-day, hour-to-hour and within an hour. Traffic volume patterns day-to-day often show remarkable similarity and these patterns are useful for prediction. Usually, traffic pattern differ between Weekdays and Weekends due to different travel demand. Within a day, traffic volume can also vary significantly. There are usually two peaks during a typical day: rush hours or peak hours, once in the morning and once in the evening. The spatial distribution of the traffic volume patterns can also be different, due to the different roadway capacities, traffic demand, and other factors. Usually, the farther apart of the

locations, the more different the traffic flow patterns of these locations. On the other

hand, traffic flow data obtained from two closely spaced detectors often show

similarities. Later sections will discuss multivariate forecasting that makes use of the

spatial correlations of traffic data.

1.1.2 Speed

Speed is a quality measurement of travel since the travelers are more concerned

about the time they spend on the road, which is related to travel speed. The definition of

instantaneous speed is

$$u_i = \frac{dx}{dt}, \tag{1}$$

where x is the length of the path traveled until time t, i represent different vehicles. In the

literature, there are several different ways of calculating the average speed of a group of

vehicles. One way is by taking the arithmetic mean of the observed data. This is termed

the time mean speed, and the equation is as below:

$$\bar{\mu}_t = \frac{1}{N}\sum_{i=1}^{N}\mu_i, \tag{2}$$

where N is the number of vehicles passing the fixed point. The other way is the space

mean speed: the total length of a roadway segment divided by the total time used to

travel this segment. The time mean speed is always greater than or equal to the space

mean speed.

1.1.3 Density

Density is the number of vehicles observed and measured over a certain road

segment. If only point detectors are available, one derives it from other variables, either

from speed and flow or from occupancy. Several equations exist to derive density from other parameters. However, most equations are only valid under certain conditions.

1.2 Short-term Traffic Flow Forecasting

There are two categories of traffic flow forecasting: long-term and short-term. Long-term traffic flow forecasting is mostly used for planning purpose. The short-term traffic flow forecasting usually finds its application in traffic operations, particularly in intelligent transportation system. Short-term traffic flow forecasting bases the predictions on using the current and the historical data to predict the traffic flow information for the next 5 to 30 minutes (Sun and Zhang 2007).

Different forecasting time intervals will have different effect on the forecasting accuracy. Usually, the forecasting accuracy improves as the time interval becomes larger. This is because the variance of traffic flow decreases as traffic flow is aggregated into longer time intervals. A study by Guo et al. (2007) felt that the establishment of the time interval for data collection is critical in determining the nature and utility of traffic flow data. In his research, various data collection time intervals were investigated. A wide spectrum of data collection time intervals from 20 seconds to 30 minutes and forecasting methods for each of these time intervals was studied. His study results indicated that the longer the data collection interval, the more stable the traffic flow data. The purpose for data use is another criterion that determines the forecasting interval. For example, if we use it in proactive signal timing design, information about future traffic flow in the next traffic circle will be critical. The HCM2000 suggests using a fifteen-

minute traffic flow rate for operational analyses. This study focuses on short-term traffic flow forecasting using a five-minute time interval.

There are two general categories of short-term traffic flow prediction methods: a univariate and a multivariate forecasting method, based on whether or not data from only one single location is used. The univariate method studies and forecast traffic flow parameters from each detector individually, while the multivariate method takes advantage of traffic flow information in nearby locations to forecast traffic flow parameters. The univariate method, when compared with the multivariate method, is more flexible and can adjust to specific traffic flow characteristics at a certain location. The multivariate method, on the other hand, can deal with the missing data by using traffic information taken from nearby sites, or those sites with similar traffic flow patterns (Kamarianakis and Prastacos 2003). Whether or not to use multivariate or univariate forecasting method depends on the traffic characteristics of the studied sites, and whether or not there is missing data. If data are obtained from several closely spaced detectors and traffic flow at these locations have similar patterns, multivariate model can be applied. If data are obtained from loosely spaced detectors, traffic flow at these locations may not have significant correlation; univariate forecasting method will perform better in this kind of situation.

2. INTRODUCTION: PROBLEM STATEMENT

This section divides existing univariate traffic flow forecasting methods into five subcategories and conducts literature reviews for each of these subcategories. It addresses two problems of existing studies—forecasting accuracy and missing data situations. It also discusses existing studies on volatility methods and multivariate forecasting methods to solve these two problems.

2.1 Existing Univariate Traffic Flow Forecasting Models

A significant number of univariate traffic flow forecasting models exist in the literature. Some of these models gained popularity among researchers and have been more thoroughly investigated. This paper divides existing model into several subcategories: Heuristic Methods, Linear Methods, Nonlinear Methods, Hybrid Methods, and Traffic Theory Methods.

2.1.1 Heuristic Methods

Heuristic methods are experience-based problem solving techniques. This kind of methods can provide a reasonable solution but not necessarily the best one in situations that an exhaustive search is impractical. Existing Heuristic methods in traffic flow forecasting area include: Random Walk (which only utilizes the current traffic information), Historical Average (predicted values are based on the average of all correspondingly observed historical traffic flow data), Informed Historical Average (the

combination of a Random Walk method and a Historical Average method) and Urban Traffic Control System predictor (UTCS)(William, B.M. 1999). Generally, the Heuristic methods are relatively easy to implement and can speed up the process of finding a good but not perfect solution. However, they do not investigate the dynamics nature of the traffic flow data and only arbitrarily unitizes the historical pattern or current value of the traffic flow data in forecasting.

2.1.2 Linear Methods

Short-term traffic flow forecasting techniques that are based on linear methods assume linear spatial and temporal relationships of traffic flow data. They assume the studied data sets are stationary. Exiting linear traffic flow forecasting methods are Univariate Box-Jenkins method, Exponential Smoothing method, Spectral Analysis, ARIMA model, and Kalman Filter method.

Ahmed and Cook (1979) investigated the application of the Box-Jenkins technique in freeway traffic flow forecasting and concluded that the ARIMA models were more accurate than moving-average, double exponential smoothing, and Trigg and Leach adaptive methods, in terms of mean absolute error, and mean squared error. Nicholson and Swam (1974) studied a short-term traffic flow forecasting method based on the spectral analysis of time series. Study results indicate that spectral analysis provides reasonable forecasting accuracy on traffic flow with periodic behavior. Davis and Nihan (1984) applied time-series methods to freeway level of service estimation. The time series method developed in their paper had the ability to detect relatively small

average changes in traffic flow characteristics (e.g. peak hour volume and lane occupancy), and thus can be related with freeway level of service.

Wild (1997) developed a pattern based short-term traffic flow forecasting methodology. The proposed model forecasts flow by dividing the system into three parts: pattern transformation, pattern classification and the choice of a suitable comparison pattern. His method is entirely empirical and does not consider theoretic relationships of traffic flow data. Williams et al. (1998) applied ARIMA and Winters' exponential smoothing models for traffic flow forecasting. The study results indicate that seasonal ARIMA models outperform the Nearest-Neighbor, the Neural Network, and the Historical Average classical models that have been previously developed. Ye et al. (2006) proposed a Scented Kalman Filter method to estimate flow speeds with single loop data. Their study results indicate that the proposed method outperforms other methods in forecasting accuracy. Okutani and Stephanedes (1984) developed two short-term traffic volume prediction models based on Kalman Filtering theory. The most recent prediction error is then taking into consideration when performing parameters estimation. In addition, by taking into account data from other links can improve the forecasting accuracy.

The linear model assumes linear relationship among traffic flow data and provides an easily understood and straightforward expression to traffic flow forecasting. However, if nonlinear relationships exist, its forecasting ability will be compromised. For example, the ARIMA model predicts future traffic flow information based on its historical traffic flow data. Its performance will be affected when handling missing

values or responding to unexpected events. Some more complex linear models, like

Kalman filtering method, require longer training time.

2.1.3 Nonlinear Methods

Nonlinear methods relax the assumption of a linear relationship among traffic

flow data, and thus can represent a nonlinear relationship in historical traffic flow data.

Some commonly used nonlinear methods in traffic flow forecasting field include

Wavelet Analysis, Neural Network, and Support Vector Machine methods.

Xiao et al. (2003) developed a fuzzy-neural network based traffic prediction

model, which uses the wavelet de-nosing method to eliminate the noise caused by

random travel conditions. His paper uses wavelet transform to analyze non-stationary

signals to obtain their trends; uses fuzzy logic to reduce the complexity of the data; and

uses neural network in increasing the accuracy of the prediction. Chen and Wang (2006)

decomposed traffic volume data into high frequency and low frequency components by

using wavelet transform and a neural network method to approximate signals by

summing up different signal components to get the final prediction results. Dougherty

(1995) conducted a literature review of Neural Network applications in traffic flow

forecasting field and identified over 40 papers published between 1990 and 1995. Smith

and Demetsky (1994) compared a back propagation neural network model with two

traditional forecasting methods: a historical data based algorithm and a time-series

model. Their study results showing that the back propagation model had considerable

potential for the application of short-term traffic flow forecasting. Ledoux (1997) first

constructed a local neural network on single signalized link and then applied it over

junctions of an urban street network. Park et al. (1998) applied a radial basis function

neural network to freeway traffic volume forecasting and compared it with the Taylor

series, exponential smoothing method (ESM), double exponential smoothing method,

and the back propagation neural network (BPN) method. Lam and Toan (2008) applied

the support vector regression method in travel time prediction. Castro-Neto et al. (2009)

developed an online-SVR method for short-term traffic flow forecasting under both

typical and atypical conditions and the study results indicate that the online-SVR method

outperforms other methods under non-recurring atypical traffic conditions.

Nonlinear forecasting methods have the ability to model nonlinear relationships

of traffic flow data. Moreover, they are more flexible in modeling time and space

relationships of traffic flow data. Most nonlinear forecasting methods have complex

model procedures, require pre-knowledge of traffic flow information, and are black box,

i.e. the underlining structure of the model is not clear to users.

2.1.4 Hybrid Methods

Voort et al. (1996) developed a hybrid method known as the KARIMA method,

for use in short-term traffic flow forecasting. A Kohonen map is used to ease the

classification problem and the forecasting results indicate that this hybrid method out

performs a single ARIMA model or a back propagation neural network model. Park

(2002) proposed a hybrid neuro-fuzzy application that first uses a fuzzy C-means (FCM)

method to classify traffic flow patterns into several clusters and then uses a radial-basis-

function (RBF) neural network to develop a forecast. The study results indicate that the

hybrid of the FCM and RBF method are promising in traffic flow forecasting. Chen and

Wang (2002) proposed a form of neuro-fuzzy systems (NFS) to forecast short-term traffic flow and it indicates that the NFS based approach is an effective method for short-term traffic flow forecasting.

2.1.5 Traffic Theory Methods

Based on the theory of "kinematic waves," Newell (1993) proposed a simplified version of kinematic waves to model highway traffic. In his theory, only two waves are studied: a forward moving wave for uncongested traffic and a backward moving wave for congested situation. Szeto et al. (2009) developed a multivariate, multistep ahead traffic flow forecasting model by using a cell transmission model and SARIMA model. The proposed model has the ability to capture traffic dynamics, queue spillback and traffic pattern seasonality. This study results indicate that the proposed model can predict real-times traffic flow in congested situation with frequent queue spillback occurrence. Guin (2004) investigated a new approach to incident detection, which is based on the assumption that current traffic conditions have the ability to indicate future traffic conditions. This approach constructed a discrete state propagation automatic incident detection model based on the theory of cell transmission model and was able to predict traffic state 20-second ahead.

2.2 Short-comings of Existing Models

As we discussed in the section 2.1, a significant number of forecasting methods exist in the literature and they involve techniques in multiple areas. However, most existing studies on univariate models have limitations in two aspects.

One of the shortcomings is that existing methods only focus on the expected value of traffic flow data in the next few minutes and assume that the variance is constant without considering the volatile nature of traffic flow data. However, according to the nature of traffic flow data, variability exists. Existing studies have not paid enough attention in traffic condition uncertainty forecasting. Here the definition of variability is the conditional standard deviation of traffic flow. The ability to capture the uncertainty of traffic flow forecasting results can give us more information on traffic conditions over the next few time steps. One example is that a sudden drop of traffic flow would occur in the congested situations; another example is that a sudden rise of traffic demands leads to the increased traffic flow volumes. Because variability is not directly observable, and its underlining features are relatively difficult to capture compared with the expected value of traffic flow data, most models can only capture the average value of traffic flow during a certain time period and cannot capture these unexpected changes which are also critical to travelers or transportation system managers.

The other shortcoming of existing studies is that some methods have limited forecasting abilities when part of the data used for forecasting is missing or erroneous. While the historical average method is often used to deal with this issue, the forecasting accuracy cannot be guaranteed.

2.3 Proposed Methodologies

The volatility model releases the assumption that the variance part of the time series model is constant. This method focuses on the modeling of dependencies among

residuals at different time steps. The volatility model provides a confidence interval for the forecasting results and is an indicator of the reliability of a predicted value. A limited number of literatures references exist in the traffic flow volatility model. Guo et al. (2007) developed a combination model based on the SARIMA and the GARCH method to determine the applicable data collection time intervals for short-term traffic condition forecasting. This paper use GARCH process to model the conditional variance. Kamarianakis et al. (2005) discussed the application of the GARCH model for representing the dynamics of traffic flow volatility and aimed at providing better confidence intervals for traffic flow forecasting results. The ARIMA-GARCH model was also introduced in other papers to forecast travel time variability (Sohn and Kim 2009; Tsekeris and Stathopoulos 2006). These studies indicate that the traditional time series method is promising in capturing the mean values of traffic flow data, while the GARCH model can predict time-varying conditional variance. Our research studies the application of ARIMA-GARCH model in the freeway traffic flow forecasting area and uses the one-step ahead forecasting method to get the expected values of the data and reliability. We also studies forecasting performances in both normal situations and missing data situations.

Missing traffic data occurs at certain times and locations due to failures in power or communication, malfunctioning devices, or observations which are obviously incorrect. The univariate forecasting models will not function well in this situation since its forecasting value is based on its own historical data. One should use information from other sources to deal with the missing data problems. Multivariate models consider

traffic flow information from other detectors and have the potential to deal with missing data. Upstream and downstream flows can have influence on traffic flow at studied site. An increased traffic volume in an upstream location may result in the increase of volume in a downstream location when there are no major access points between these two locations. Even if there are access points between the two locations, a correlation of flow may also exist between these two locations. A multivariate model takes into consideration correlation of flows among different locations. It can also potentially improve model performance when missing data exists for one detector.

In recent years, interests have risen in multivariate traffic flow forecasting as traffic flow information in road networks become more readily available. Chang et al. (2000) utilized data from adjacent roads while performing traffic flow forecasting, but the information of the adjacent road still was not used to its full potential. Yin et al. (2002) forecasted the downstream flow by utilizing upstream flows in the current time interval and chose a fuzzy-neural model as the forecasting methodology. Pfeifer and Deutsch (1980) studied the multivariate method, predicted traffic parameters in a road network, and used the space-time autoregressive integrated moving average model to forecast. Kamarianakis and Prastacos (2003) applied the STARIMA methodology to represent traffic flow patterns in an urban network. In their research, the STARIMA model incorporated spatial characteristics by using weighted matrices, which were estimated based on the distances between data collection points. Jin and Sun (2008) applied multitask learning (MTL)-based neural networks to urban vehicular traffic flow forecasting. The authors incorporated traffic flows at different locations into the input

layer of the back propagation (BP) neural network. Although the study results show that the MTL in BP neural networks are promising and are effective approaches for traffic flow forecasting, they do not consider the spatial correlations. Sun and Zhang (2007) modeled traffic flows along adjacent road links in a transportation network similar to a Bayesian Network.

This research uses the VAR model and the GRNN model to perform traffic flow forecasting in missing data situations. The VAR model is an extension of an ARIMA model in the multivariate analysis field. The VAR model use historical traffic flow data obtained from two closely spaced detectors to forecast future flow information at these two detectors. The assumption of the GRNN model is that the upstream freeway traffic flow data can provide adequate information to forecast of down-stream traffic flow. If large percentages data missing from a certain detector, traffic flow information from its closest up-stream detector is used as the model input to predict next time step traffic information at the point of interest.

3. UNIVARIATE TRAFFIC FLOW FORECASTING

Traffic flow data is predominantly collected at detectors, such as inductive loop detectors (ILDs), microwave detectors, and video detectors at certain points. These kinds of data collection technologies are capable of providing volume counts and speed data during a specified time period. If properly installed and maintained, one can obtain historical and current traffic flow data from these devices. Thus, the traffic flow data can provide real time traffic flow information for road users and managers. While the current traffic information is important, the information arrives too late for the purpose of proactively managing and coordinating the control of traffic. Knowledge of the near future traffic information is critical for proactive control systems. Because a traffic stream varies over both time and space, traffic flow data detected at different times and different locations are parameters of statistical distribution: not absolute numbers (Lieu 1999). This section proposes a univariate traffic flow forecasting method to capture the time variance of the traffic information. The univariate method studies and forecasts traffic flow parameters at each detector individually, without considering the spatial correlation of traffic parameters.

As discussed section 2.1, a significant amount of univariate short-term traffic flow forecasting methods exists in the literature. This kind of forecasting methodology use both historical and current traffic flow data obtained at the point of interest to predict the future roadway conditions. One limitation of existing single-point forecasting

methods is that most methodologies only focus on the expected value of traffic-flow data and ignore the volatile nature of the traffic stream. Traffic flow varies significantly for near congested situations. However, only forecasting the expected value of traffic parameters cannot provide adequate information. Accurate predictions of the variance part can indicate whether or not there will be a big change in traffic flow over the next few minutes. In addition, by making use of other traffic information (like speed), we can also decide if there will be a big drop or increase of traffic flow in the near future; thus forecast the traffic condition in the next time step.

This section presents the application of volatility models in single-point traffic-flow forecasting. The purpose of this model is to predict the shift of traffic conditions based on historical traffic flow data. The basic idea of a volatility model is to first fit the expected values of the data set and then assign a volatility model to study the variance part. Because existing study results indicate that the ARIMA model provides adequate forecasting results for the traffic flow data, we will use the ARIMA model to forecast the expected values of the traffic flow data and use the GARCH model to study the variance part. The rest of this section introduces theoretical background of ARIMA model, which includes order selection, parameters estimation, and data transformation. Then it covers the basic concept of a volatility model and two classical volatility models: the ARCH and GARCH models.

3.1 Autoregressive Integrated Moving-Average (ARIMA) Model

ARIMA models are one of the most general classes of time series models, in which data can be made stationary by transformations such as differencing and logging. A non-seasonal ARIMA model is classified as an ARIMA (p,d,q) model, in which: p is the number of autoregressive terms, d is the number of non-seasonal differences and q is the number of lagged forecast errors (Nau 2005). An understanding of three common processes is prerequisite to understand the autoregressive integrated moving average (ARIMA) process better. These three processes are autoregressive (AR) model, moving average (MA) model, and autoregressive moving average (ARMA) model.

3.1.1 Three Common Processes

3.1.1.1 Autoregressive Model

The basic idea of an autoregressive model is that the current value in the time series is a function of its past values. Assume we have a time series dataset$\{x_t\}$, the value of x_t can be represented by its p past values$\{x_{t-1}, x_{t-2}, \ldots, x_{t-p}\}$. By looking at the autocorrelation function, one can assess the order of p.

Equation representation of an autoregressive model of order p, abbreviated AR (p) is as follows:

$$x_t = \emptyset_1 x_{t-1} + \emptyset_2 x_{t-2} + \cdots + \emptyset_p x_{t-p} + \omega_t \qquad (3)$$

where p is the extension of past values used for prediction, x_t is stationary, $\emptyset_1, \emptyset_2, \ldots, \emptyset_p$ are constants ($\emptyset_p \neq 0$), and ω_t is a Gaussian white noise series with mean zero and variance σ_w^2.

If using the backshift operator B, the equation becomes as

$$(1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p)x_t = w_t \tag{4}$$

Here the definition of backshift operator is

$$B^k x_t = x_{t-k} \tag{5}$$

Let $\emptyset(B) = 1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p$, the equation can be expressed more concisely as

$$\emptyset(B)x_t = w_t \tag{6}$$

3.1.1.2 Moving Average Model

The moving average model assumes x_t is a linear combination of white noise ω_t. The definition of moving average model of order q is as

$$x_t = \omega_t + \theta_1 \omega_{t-1} + \theta_2 \omega_{t-2} + \cdots + \theta_q \omega_{t-q} \tag{7}$$

where q is lags that are used for the prediction of x_t, x_t is stationary, $\theta_1, \theta_2, \dots, \theta_p$ are constants ($\theta_p \neq 0$) and ω_t is a Gaussian white noise series with mean zero and variance σ_w^2.

If we use the backshift operator, and let $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_p B^p$ then

$$x_t = \theta(B)\omega_t \tag{8}$$

3.1.1.3 Autoregressive Moving Average Model

Another important parametric family of the time series is the autoregressive moving-average, or ARMA, processes. The mathematical representation of ARMA (p,q) process is as follows:

$$x_t - \emptyset_1 x_{t-1} - \emptyset_2 x_{t-2} - \cdots - \emptyset_p x_{t-p} = \omega_t + \theta_1 \omega_{t-1} + \theta_2 \omega_{t-2} + \cdots + \theta_q \omega_{t-q} \ (9)$$

in which $\{x_t\}$ is stationary, ω_t is a Gaussian white noise series with mean zero and variance σ_w^2, and the polynomials $(1 - \phi_1 z - \cdots - \phi_p z^p)$ and $(1 + \theta_1 z - \cdots - \theta_q z^q)$ have no common factors.

Let $\emptyset(z) = 1 - \emptyset_1 z - \emptyset_2 z^2 - \cdots - \emptyset_p z^p$ and $\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_p z^q$

The more concise representation of the equation is

$$\phi(B)X_t = \theta(B)Z_t \tag{10}$$

The upper equation indicates that if $\theta(z) \equiv 1$, the time series is an autoregressive process of order p, and it is a moving-average process of order q if $\phi(z) \equiv 1$.

If the data does not exhibit apparent deviation from stationary and its autocovariance function decreasing rapidly, then we can fit an ARMA model to this data. If the data does not follow the previous two properties, we can try a transformation of the data, which generates a new time series that process the two properties. One of the most commonly used transformations is differencing, which leads to the concept of the ARIMA model.

3.1.2 Autoregressive Integrated Moving-average Model

The ARIMA model is a generalization of an autoregressive moving average (ARMA) model. This model is applied in cases that the original time series data does not show evidence of a ARMA model, but proper transformation (corresponding to the "integrated" part of the model) of the original data can fit an ARMA model.

The ARIMA (p,d,q) model of the time series $\{x_1, x_2, ...\}$ is defined as

$$\phi(B)(1 - B)^d X_t = \theta(B)Z_t, \{Z_t\} \sim WN(0, \sigma^2), \tag{11}$$

in which $\phi(z)$ and $\theta(z)$ are polynomials of degrees p and q respectively, and $\phi(z) \neq 0$ for $|z| \leq 1$.

Some special cases of ARIMA model are ARIMA(0,1,0) - random walk, ARIMA(1,1,0) - differenced first-order autoregressive model, ARIMA(0,1,1) - simple exponential smoothing model. To identify an appropriate ARIMA model for the studied time series data, the first step is finding an appropriate transformation for the data that can fit a ARMA (p,q) model. The second step is to decide the order of the ARMA (p , q) model, and the last step is parameter estimation.

3.1.2.1 Transformation Technology

The first step of time series analysis is to plot the original data. The classical decomposition model indicates that a time series data can be decomposed as a trend component, seasonal component, and a random noise component. A cursory look at the plot of the original data is needed to check whether or not there is an obvious trend or seasonal component in the data sets. In a time series analysis, we need to remove the

trend and seasonal component, if there is any, to get stationary residuals. A preliminary

transformation of the data can help to achieve the goal.

Transformation of original data is one of the most commonly used technologies

in trend and seasonal components removing. Box and Jenkins (1970) developed an

approach by applying a differencing operator repeatedly to the original time series data

until it resembles a realization of some stationary time series $\{W_t\}$. Then we can use the

theory of the stationary time series $\{W_t\}$ to model and forecast the $\{W_t\}$ series and hence

the original time series. In the ARIMA (p,d,q) model fitting process, if the original data

set shows a slowly decaying positive sample autocorrelation function, we would

naturally apply the operator $\nabla = 1 - B$ repeatedly until the autocorrelation function show

rapidly decaying feature. In this model, d represents the number of differencing of

original data set.

3.1.2.2 Order Selection and Parameter Estimation

After proper transformation, the next step is to select the appropriate order p and

q for the ARMA model. It is not a wise choice to select p and q arbitrarily large from a

forecasting point of view. To avoid over-fitting problems, penalty factors is introduced

to discourage the fitting of models with too many parameters. Some widely used criteria

for model selection are FPE, AIC, and BIC criteria of Akaike and AICC. The best model

is selected based on the smallest value of one for these criteria.

The R Language uses two methods for parameter estimation: maximum

likelihood and minimize conditional sum-of-squares. If there are no missing values in

the original data, the default method is to use conditional-sum-of-squares to find starting

values, then using maximum likelihood to find the optimal parameters.

3.2 Volatility Models

In statistics, heteroscedastic, or heteroskedastic is a sequence of random variables

that has different variances. Traditional forecasting methods assume constant variance of

the data when perform forecasting. If heteroscedastic exist, the crucial question is the

prediction accuracy of the model. In this case, the critical issue is to model the variance

part of the error terms and then to find out what makes them large.

A cursory look at traffic flow data indicates that the variances of traffic flow data

over some time periods are greater than that at other time periods. A volatility measure-

like a standard deviation- can be used in accident, congestion, and abnormal situations.

While many specifications only consider the expected value of traffic flow data and have

been used in traffic flow forecasting, virtually no methods have been used for the

variance forecasting before the conditional heteroscedastic models were introduced.

Some time series data is serially uncorrelated but dependent. The basic idea of the

volatility models is to capture the dependency in this kind of time series data. The

structure of the model can be writen as the sum of the mean and the variance:

$$y_t = u_t + e_t, \tag{12}$$

where y_t is the observed data at time t, here it represents traffic flow data at time t,

$u_t = E(y_t | F_{t-1})$ is the conditional mean of y_t, F_{t-1} , which denotes the information set

available at time t-1 and e_t is the variance of y_t.

Most existing prediction models concentrate on the conditional mean part and assume that the variance part simply satisfies the white noise properties. Although this assumption simplifies the structure of the fitted model, the prediction accuracy can be compromised. Conditional heteroscedastic models relax the assumption and treat e_t as conditionally heteroscedastic. So the following expression of e_t is:

$$e_t = z_t\sqrt{h_t} ,$$
(13)

where z_t is independent and identically distributed with zero mean and unit variance and $h_t = E(e_t^2|F_{t-1})$.

Equation (13) indicates that the conditional distribution of e_t is independent and identically distributed with zero mean and variance of h_t. Volatility models are concerned with time-evolution of the conditional variance of traffic flow data. Different ways to address the conditional variance of e_t leads to different heteroscedastic models.

3.2.1 AutoRegressive Conditional Heteroskedasticity (ARCH) Model

In 1982, Engle proposed the ARCH model, which is the first model that provides systematic framework for volatility analysis. The basic idea of the ARCH model is that the conditional variance is a linear combination of past sample variance. An ARCH (q) model assumes that

$$e_t = z_t\sqrt{h_t}$$
(14)

$$h_t = a_0 + \sum_{i=1}^{q} a_i e_{t-1}^2,$$
(15)

where $\{z_t\}$ is a sequence of independent and identically distributed random variables with mean zero and variance 1, which often assumes to follow a standard normal,

standard student-t or generalized error distribution. The structure of the model indicates

that large past sample variance leads to large conditional variance for the innovation e_t,

which further indicates that larger past value of sample variance tends to be followed by

another large sample variance. In other words, if the past value of variance is large, the

probability of obtaining a large variance is greater than that of obtaining a small

variance.

3.2.2 Generalized Autoregressive Conditional Heteroskedasticity (GARCH) Model

Although the structure of the Autoregressive Conditional Heteroskedastic

(ARCH) model is simple and easy to understand, many parameters are often required to

adequately describe the volatility process of a time series. Bollerslev (1986) proposed a

useful extension known as Generalized Autoregressive Conditional Heteroskedasticity

(GARCH) model based on the idea of the ARCH process, allowing for a much more

flexible lag structure. The idea of the extension of the ARCH process in the GARCH

model resembles the extension of the AR process in the ARMA process.

$$e_t = z_t \sqrt{h_t}, \tag{16}$$

$$h_t = a_0 + \sum_{i=1}^{q} a_i e_{t-1}^2 + \sum_{i=1}^{p} \beta_i h_{t-i}, \tag{17}$$

in which $p \geq 0, \ q > 0, a_0 > 0, a_i \geq 0, i = 1, \dots, q_, \beta_i \geq 0, i = 1, \dots, p_,$ and

$\sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) < 1.$

For $p = 0$, the process becomes the ARCH (q) process, and for p=q=0, the

process becomes white noise. The difference between the GARCH and the ARCH

process is that the GARCH (p, q) process not only has past model sample variances but

also has lagged conditional variance, as well.

4. MULTIVARIATE TRAFFIC FLOW FORECASTING METHOD

A univariate model only considers historical traffic flow data from a single point. In other words, the current or future value of the data set is explained only by its own past or current values. Although the univariate forecasting results indicate that historical traffic flow data can provide adequate information for future traffic condition, the forecasting accuracy will be affected when part of the historical data is missing for that particular detector. In this situation, one should consider other influential factors to improve the forecasting accuracy. The multivariate forecasting methods consider traffic information in upstream locations when performing forecasting. Two different methods are used: one is the Vector Autoregression (VAR) model and the other is General Regression Neural Network (GRNN). This section studies the forecasting performance of these two models by considering upstream information.

According to traffic flow theory (Lieu 1999), locations are important to the study of traffic variables. A simple example that explains the influence of locations at closely spaced segments of roadway is shown in Figure 1.The simple representation of the speed-flow curve will be used to illustrate the problem. One assumption is made in this example, the underlining speed-flow curves of these three locations are the same. To not oversimplify the problem, a major entrance and an exit ramp are added to between locations A and B and, B and C. The entrance ramp will add a considerable flow to location B and the exit ramp will remove a significant portion of traffic flow at location

B. So traffic demand at location B will be the highest. When location B reaches its

capacity, traffic flow at this point reach the highest value (location A in Figure 1.). A

queue will back up towards upstream traffic, location A will have a stop and go traffic

(point B in Figure1.). At the same time, because there is an exit ramp between location B

and C, traffic demand at location C will less than that at location B and it will not reach

its capacity. This example indicates that locations are important to traffic flow characters

in different road segments. In this example, if all locations do not reach their capacity,

their flow at these three road segments should all at its upper part of the curve. If

Location B reaches its capacity and results in back-up effect to location A, flow rate at

location A will decrease and flow rate at location B will reach its highest value.
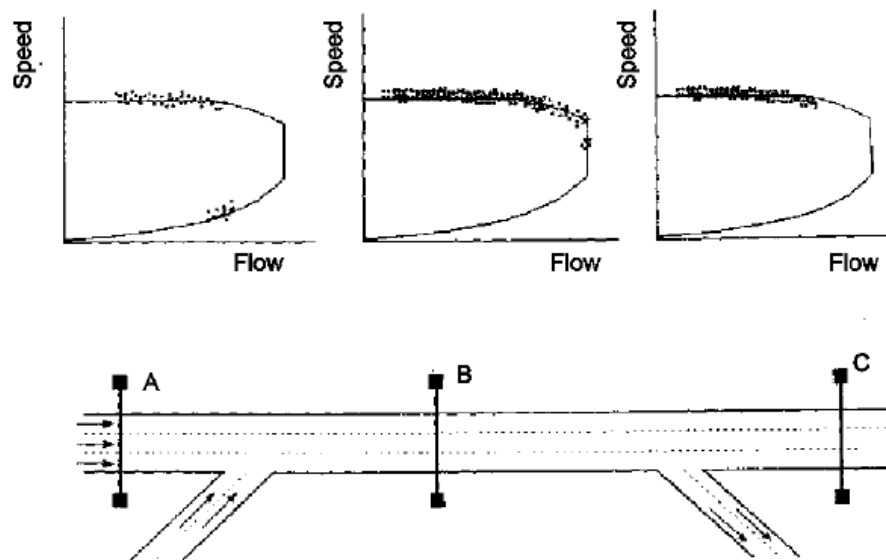
Figure 1. Effect of locations to traffic flow patterns (Lieu 1999)

Since traffic flow will be influenced by its upstream and downstream flow rates, one should consider its upstream and downstream information when study flow characteristics at point of interest. As for prediction purposes, upstream traffic flow at time t will, to a certain extent, influence downstream traffic in the near future. Considering the upstream traffic information may lead to more accurate forecasting results. Driven by this, we proposed two multivariate traffic flow forecasting methods in the following subsections.

4.1 Vector Autoregression (VAR) Model

An ARIMA model only study past values of one time series. The vector autoregression (VAR) model is an extension of the univariate autoregressive model and it is one of the most successful and easy to understand models in the multivariate time series analysis. The VAR model explains a studied time series not only based on its own past values but also based on other variables. It is proven to be an efficient multivariate forecasting model in economic and financial time series field. It is also a flexible model that can represent the correlation of multiple time series.

Before conducting the analysis of relationships between two times series data, a test for unit roots is needed. If the two studied time series models have unit root, we need to figure out whether or not there is a common stochastic trend in these two models. The augmented Dickey–Fuller test (ADF) developed by Said and Dickey (1984) provides a general approach for unit root testing. The null hypothesis of this test is unit root exists

in the model. So the smaller the p-value is, the more unlikely there is a unit root in the model.

An autoregressive model can be treated as a simple regressor on several time series variables and it can capture the evolution and the interdependencies between multiple time series. Its matrix notation is:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{k,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} + \begin{bmatrix} a_{1,1}^1 & \cdots & a_{1,k}^1 \\ \vdots & \ddots & \vdots \\ a_{k,1}^1 & \cdots & a_{k,k}^1 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \vdots \\ y_{k,t-1} \end{bmatrix} + \cdots + \begin{bmatrix} a_{1,1}^p & \cdots & a_{1,k}^p \\ \vdots & \ddots & \vdots \\ a_{k,1}^p & \cdots & a_{k,k}^p \end{bmatrix} \begin{bmatrix} y_{1,t-p} \\ y_{2,t-p} \\ \vdots \\ y_{k,t-p} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \\ \vdots \\ e_{k,t} \end{bmatrix}$$

The above equation is a VAR(p) model that involves k variables. In the equation, c is the intercept, a is the parameter for the model and e is white noise.

This study focuses on the bivariate vector autoregressive model with two dependent time series $y_{1,t}$ and $y_{2,t}$. The simplest form of the VAR model is VAR (1), in which only two explanatory variables are included: $y_{1,t-1}$ and $y_{2,t-1}$. The equations of the VAR (1) model are:

$$y_{1,t} = a_{11}y_{1,t-1} + a_{12}y_{2,t-1} + e_{1,t} \tag{18}$$

$$y_{2,t} = a_{21}y_{1,t-1} + a_{22}y_{2,t-1} + e_{2,t} \tag{19}$$

There are two assumptions in the error term: the expected values of the residuals are zero and the two error terms are uncorrelated. In the vector autoregression model, expected traffic flow information at studied location is a linear combination of historical traffic flow data from the upstream location and the studied location.

The order selection of the VAR model is a trade-off between the forecasting accuracy and abbreviate of the model. If the lag length is too short, the equation cannot

provide adequate information and can lead to inefficient estimators. On the other hand, the degree of freedom will decrease with an increasing number of parameters, which also will lead to inefficient estimators. Determination of the lag length of the VAR model can be obtained from the autocorrelation plot or based on the smallest AIC.
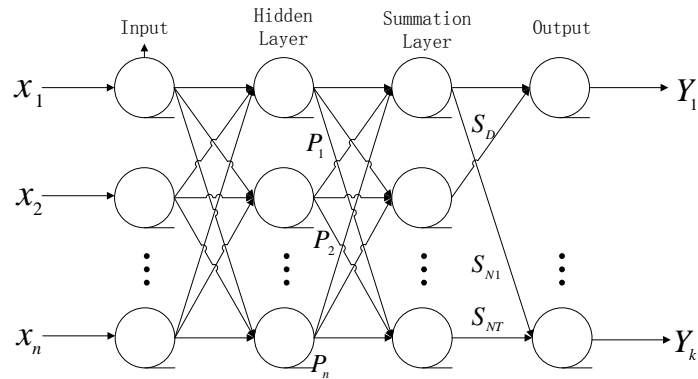
4.2 General Regression Neural Network (GRNN) Model

GRNN is derived from the RBF neural network. Its theoretical background is general regression analysis. The formula for the regression is shown in equation (20).

$$\hat{Y}(X) = \frac{\sum_{i=1}^{n} Y_i \exp[-\frac{(X - X_i)^T (X - X_i)}{2\sigma^2}]}{\sum_{i=1}^{n} \exp[-\frac{(X - X_i)^T (X - X_i)}{2\sigma^2}]} \tag{20}$$

Equation (20) indicates that the estimation of output $\hat{Y}$ given input X is the weighted average of each training sample Y_i. Each Y_i is weighted according to the exponential value of its Euclidean distance from X. Normal distribution is used as the probability function and the mean is each training sample X_i. The standard deviation or the smoothness parameter σ is subject to the searching process. GRNN can solve nonlinear problems without having to estimate many parameters, and its training time is shorter compared with other BP methods. Thus, GRNN is used as forecasting methodology in this study.

The structure of the GRNN is shown in Figure 2. The GRNN model has four layers: input, hidden, summation and output. Functions of each layer are introduced below:

Figure 2. Structure of general regression neural network

Input Layer: The number of neurons in the input layer is equal to the number of predictor variables and each X_i represents a predictor variable. The function of input layer is to standardize the range of the values so that it ranges from -1 to 1, and feed standardized values to the second layer-hidden layer.

Hidden (Pattern) Layer: The hidden layer computes the exponential value of the squared Euclidean distance between predictor variable X and training sample X_i. Then the result p_i is forwarded to the summation layer.

$$p_i = \exp[-\frac{(X - X_i)^T (X - X_i)}{2\sigma^2}] \qquad (i = 1,2,...,n)$$

$$(21)$$

Summation Layer: There are two kinds of neurons in summation layer: One kind of neuron is a denominator summation unit and it is the denominator of equation (20). It

adds up the values that come from each of the hidden layers. Equation (22) represents

the denominator summation unit.

$$s_d = \sum_{i=1}^{n} p_i$$

(22)

The other kind of neuron is the numerator summation unit and it is the numerator of

equation (20). It also adds up the weighted values that come from each of the hidden

layers. The weight for the ith neuron in the pattern layer and the jth neuron in

summation layer is y_{ij}. Equation (23) is the representation of the numerator summation

unit.

$$S_{Nj} = \sum_{i=1}^{n} y_{ij} p_i \qquad j = 1,2,...,k$$

(23)

Output Layer: The output layer divides the numerator summation unit S_{Nj} by

denominator summation unit S_d and use it as the value of the predicted target.

5. DATA DESCRIPTION AND APPLICATION

This study uses traffic data collected from six radar sensors located on U.S. Highway 290 (or U.S. 290) to conduct model fitting and forecasting. U.S. 290 is an east-west U.S. Highway located within the state of Texas. The studied segment (Northwest Fwy) begins at Sam Houston Tollway and ends at the junction of Farm to Market Road 1960(FM1960) and U.S.290. Figure (3) shows the locations of the six detector sites. Since the traffic flow is directional, we use data from northwest bound direction for model training and forecasting. The IDs of the detectors from Southeast to Northwest are 1090, 3441, 3878, 2782, 3935, and 3998. Measurements take place every 30 seconds and collected information includes volume, speed, and occupancy.

Traffic flow data from January 1, 2008 to February 5, 2008 are used and have been aggregated into five minutes data points. For each day, there are 288 data points, thus the total number of data points used is 10,368. For the purpose of model comparison, we choose the 288 data points obtained from February 5, 2008 for model prediction. Table (1) shows detailed information about the data collected from these detectors.
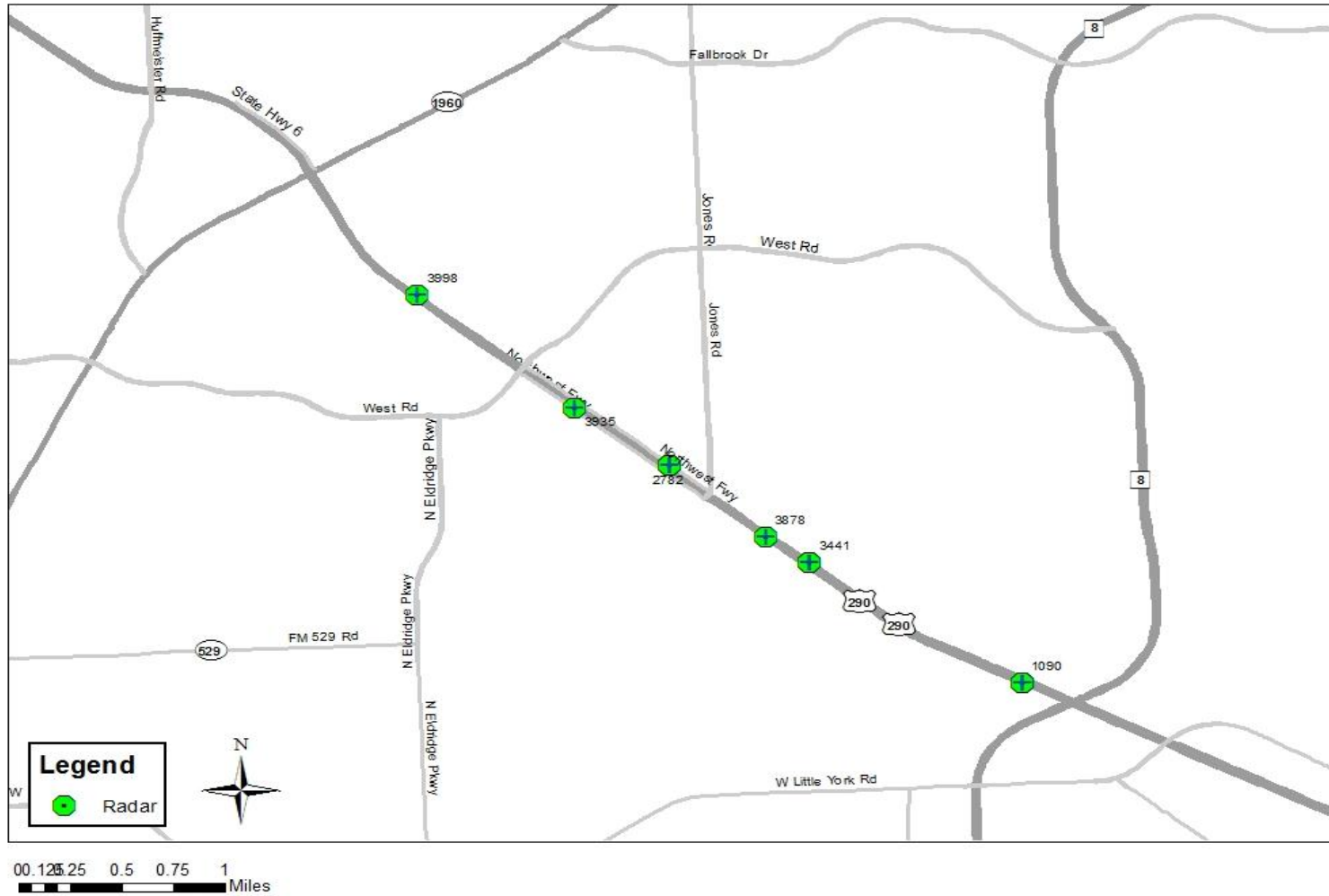
Radar Detector Locations

Figure 3. Radar detector locations for sites of interest

**Table 1** Information of six detectors at studied sites

Detector Information					
Detector ID	Name	Data Interval	# Lanes	# WB Lanes	Distance to the Next Detector
1090	US-290 Northwest@Senate IB	5 min	4	1	1.2252 Miles
3441	US-290 Northwest@FM-529 OB	5 min	7	4	0.2545 Miles
3878	US-290 Northwest@Jones IB	5 min	6	3	0.6050 Miles
2782	US-290 Northwest@Jones OB	5 min	7	4	0.5554 Miles
3935	US-290 Northwest@West IB	5 min	6	3	0.9760 Miles
3998	US-290 Northwest@Eldridge IB	5 min	6	3	

Before analyzing traffic flow data, we need a cursory look at a plot of the

original data. Based on empirical experience, traffic flow data show strong periodic

features and comparing traffic volume patterns day to day indicates remarkable

similarity. In order to give us a general idea of what daily traffic flow data looks like, we

choose to plot five-day traffic flow data from February 1, 2008 to February5, 2008..

Figure 4. and Figure 5. are five-day traffic flow data obtained from detector 1090,

detector 3441, detector 3878, detector 2782, detector 3935, and from detector 3998.
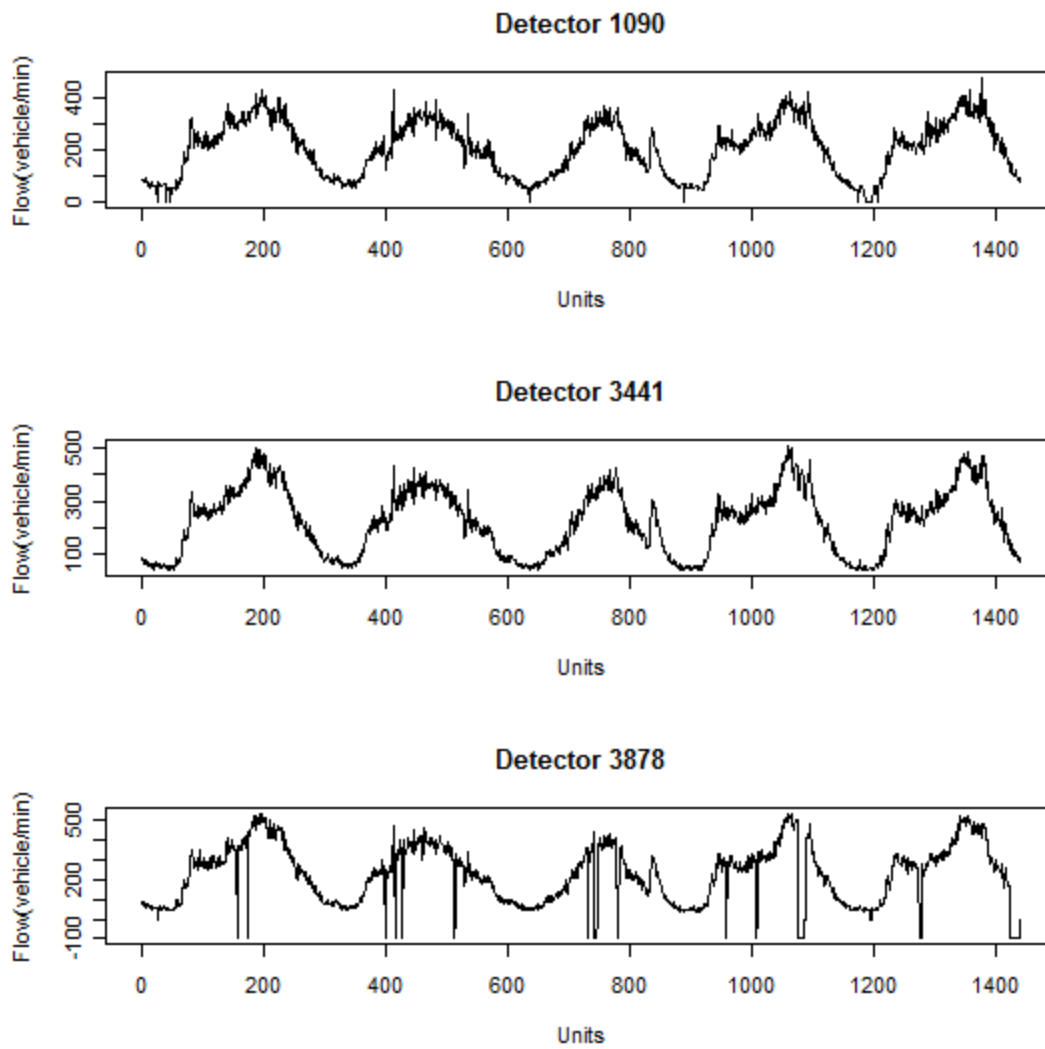
Figure 4. Traffic Flow Data from February 1, 2008 till February5, 2008(1)
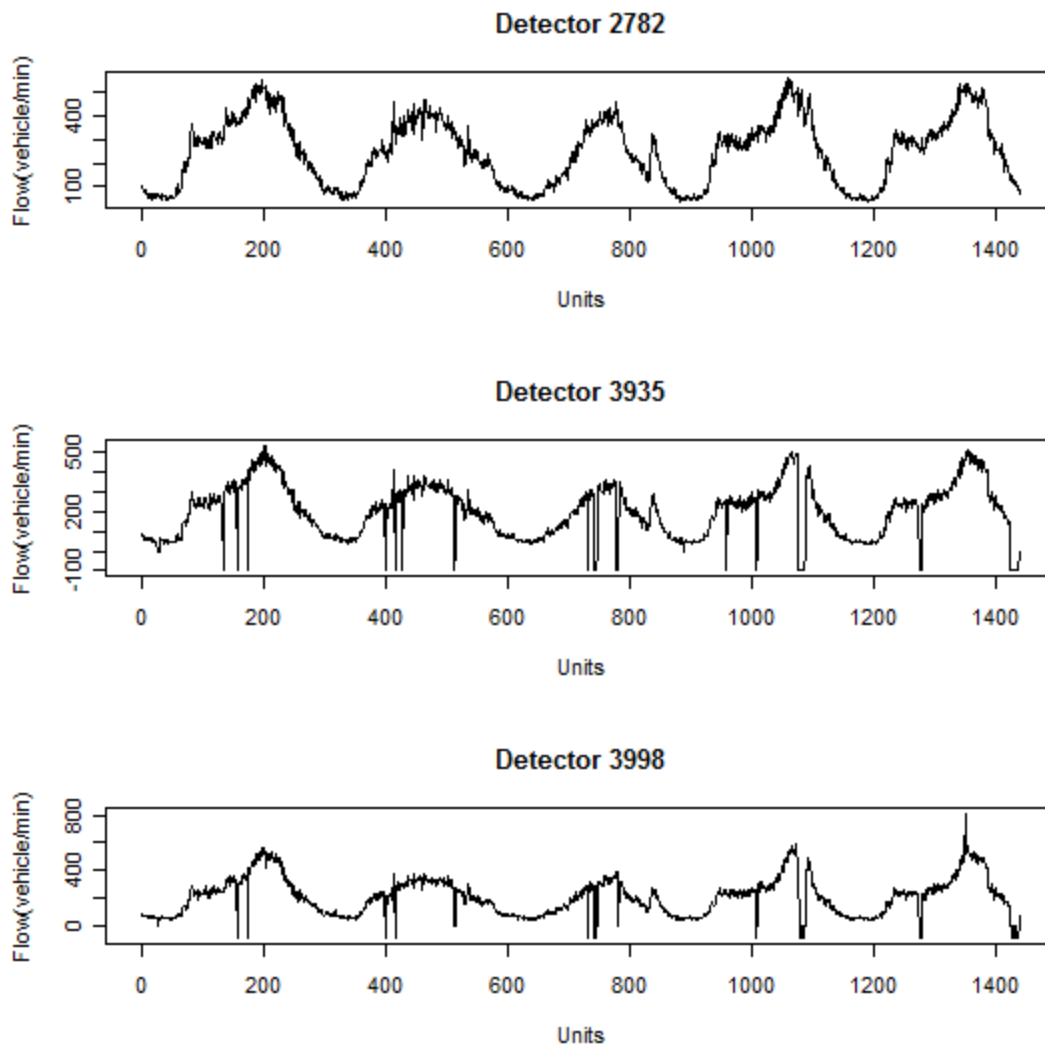
Figure 5. Traffic Flow Data from February 1, 2008 till February5, 2008(2)

Close inspection of five-day traffic flow data from these six detectors indicates that there are missing data samples from detectors 3878, 3935, and 3998. For the entire study data set, there are 3.12% missing data from detector 1090, 0.28% missing data from detector 3441 and 0.087% missing data from detector 2782. For detectors 3878, 3935 and 3998, there was more missing data. The percentages of missing data are

31.04%, 8.76% and 7.06%, respectively. For the study's dataset, if there is no data obtained during a certain time period, the traffic flow value for that period is either -1 or -99. As shown in Figure 2, there are some points that have values go below to zero which cannot be in real case. It was therefore necessary to find a proper strategy to identify the missing data.

5.1 ARIMA-GARCH Model Fitting

5.1.1 ARIMA Model Fitting

Because an ARIMA model requires relatively small number of sample data, for the ARIMA model fitting process, we use the first 4 day flow data for model training and apply the one step forecasting method for the prediction of the fifth day's traffic flow data.

We first plotted the ACF and PACF of traffic flow obtained from each detector as shown in Figure 6. and Figure 7. Although there are some differences for each plot of ACF and PACF values, they show common features: the ACF plots of all traffic flow data indicate that the auto-correlation function for the original data decreasing slowly. Common practice is to transform the original data to get a lower order model, which we are more familiar with. Further steps are needed in this case.
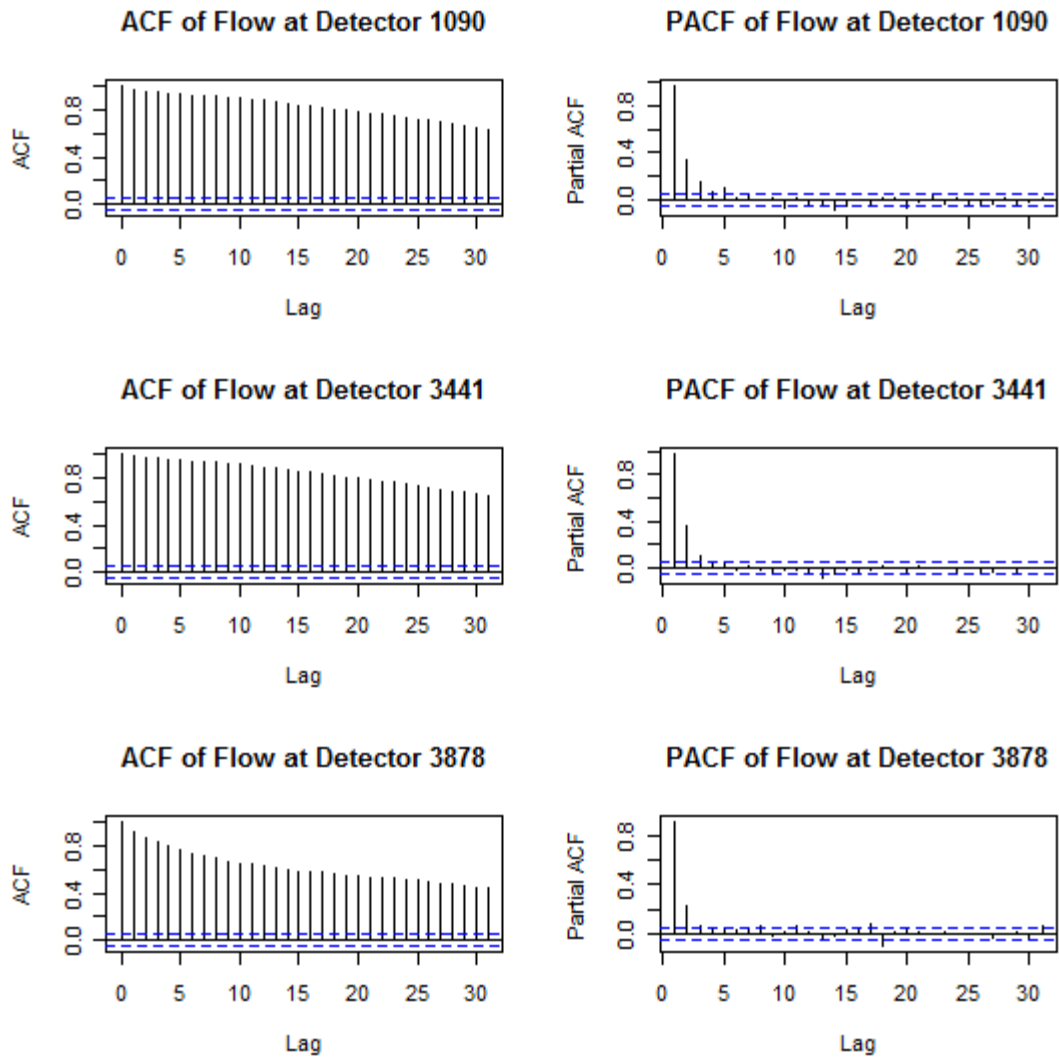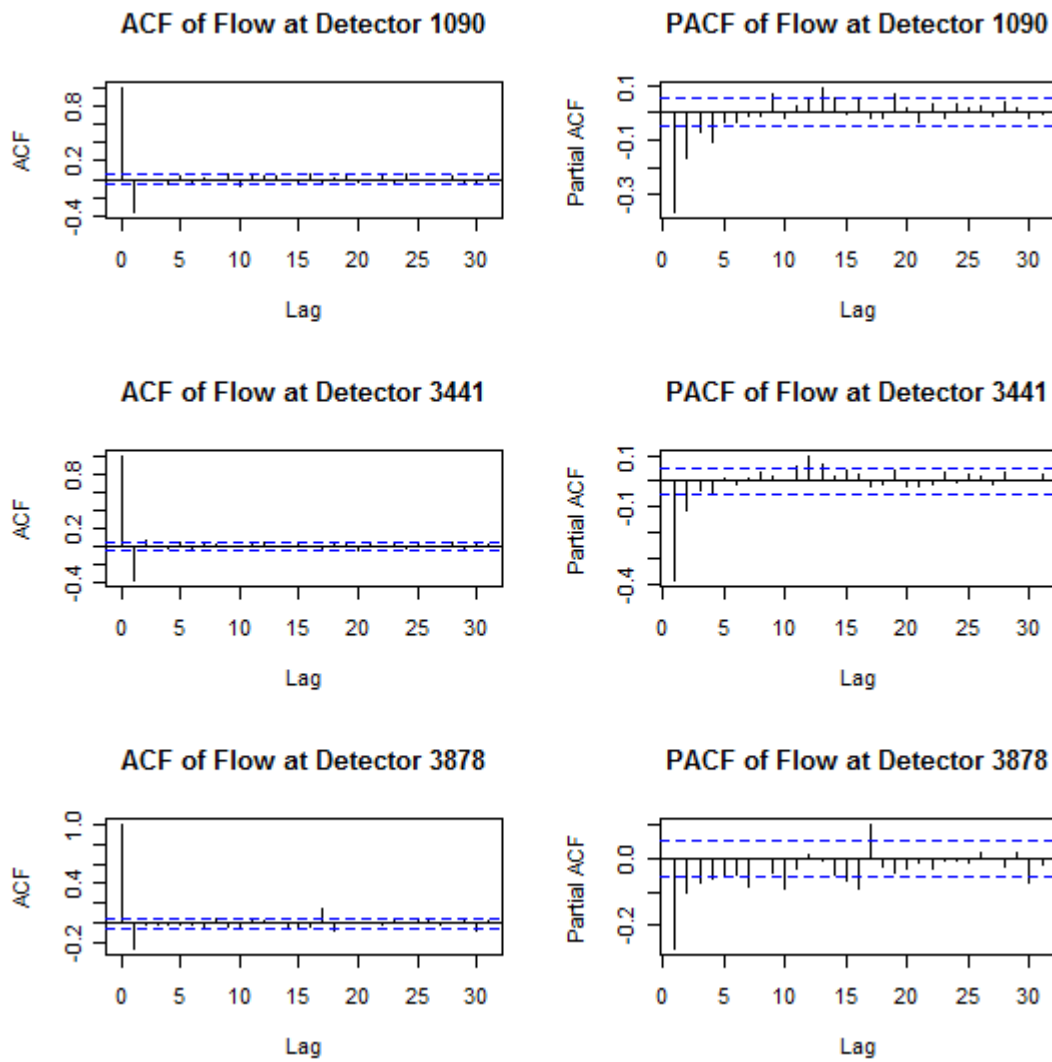
Figure 6. ACF and PACF plots of traffic flow data(1)
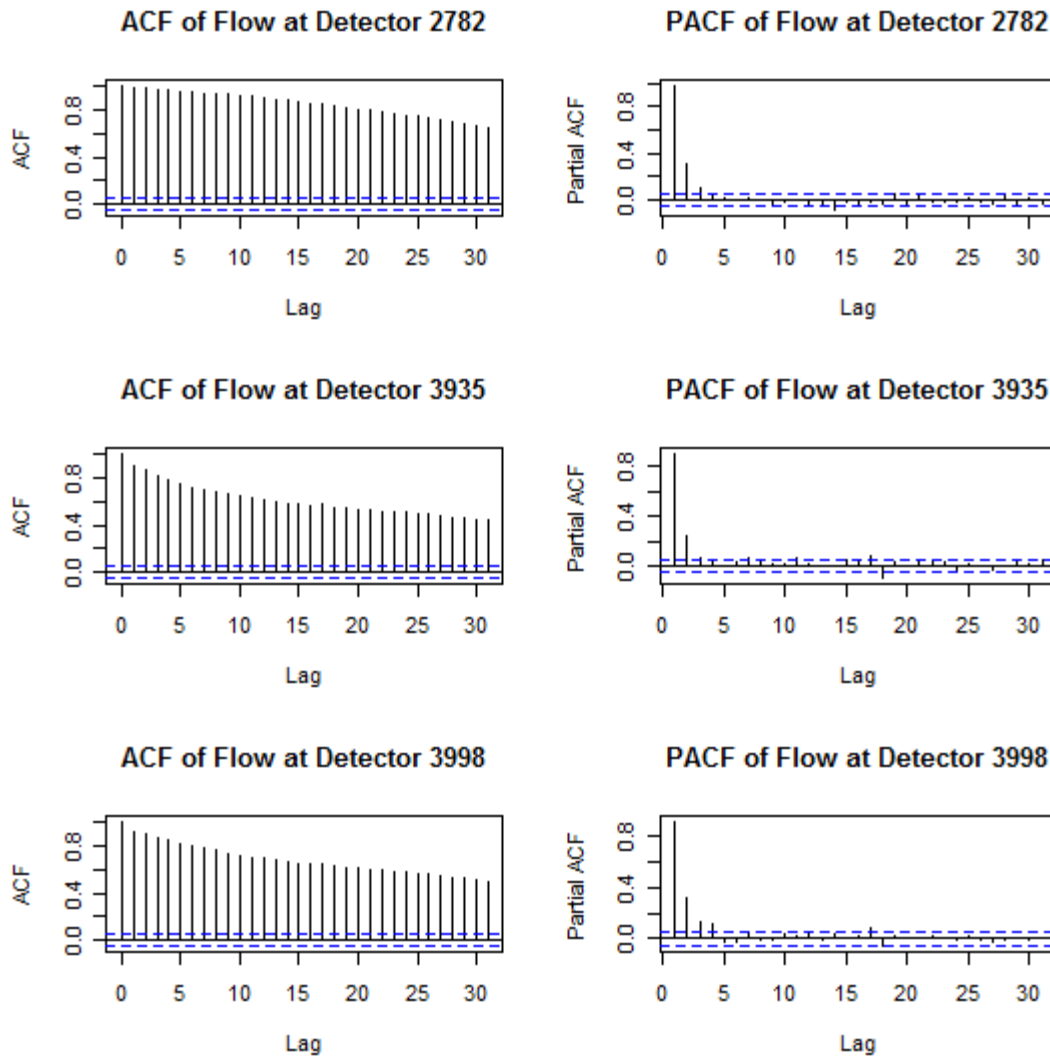
Figure 7. ACF and PACF plots of traffic flow data(2)

If the original dataset show a slowly decaying positive sample autocorrelation function, one should apply the differencing operator repeatedly until the autocorrelation function shows a rapidly decaying feature. Figure 8. and Figure9. are the ACF and PACF plot of differenced flow data at lag 1.The plots indicate that the differenced flow can be a MA(1) model since the ACF is zero except for lag 1.

Figure 8. ACF and PACF plots of differenced traffic flow data (1)
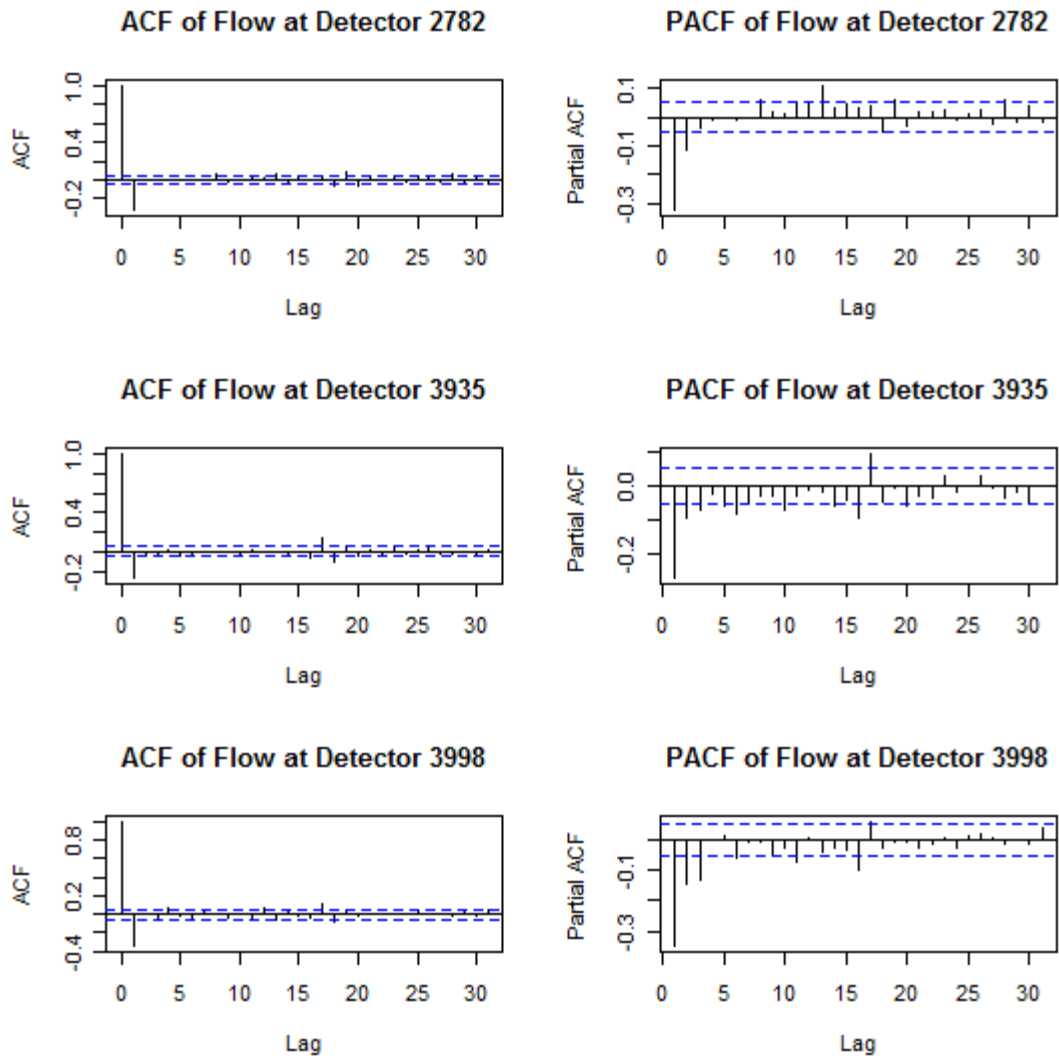
Figure 9. ACF and PACF plots of differenced traffic flow data (2)

As seen in Section 3, if we let X_t and $Y_t \equiv (I - B)X_t$ be flow and differenced flow at time t, respectively, then we can set the model as the followings:

$$Y_t \equiv (1 - B)X_t = Z_t + \theta Z_{t-1}, where\ Z_t \sim WN(0, \sigma^2)$$

The original flow X_t is a ARIMA(0,1,1) model, in which the AR order is 0, the degree of differencing is 1, and the MA order is 1. Parameter θ is the only parameter that needs to be estimated when fitting an ARIMA(0,1,1) model. The one step forecasting method is used in model prediction. A least squared error is used to fit the parameters of the model. The predicted value x_t is based on its previous values $\{x_{t-1} \dots x_{t-288}\}$; 288 data points are used for model fitting. Figure 10 and Figure 11 are plots of original data and forecasting results:
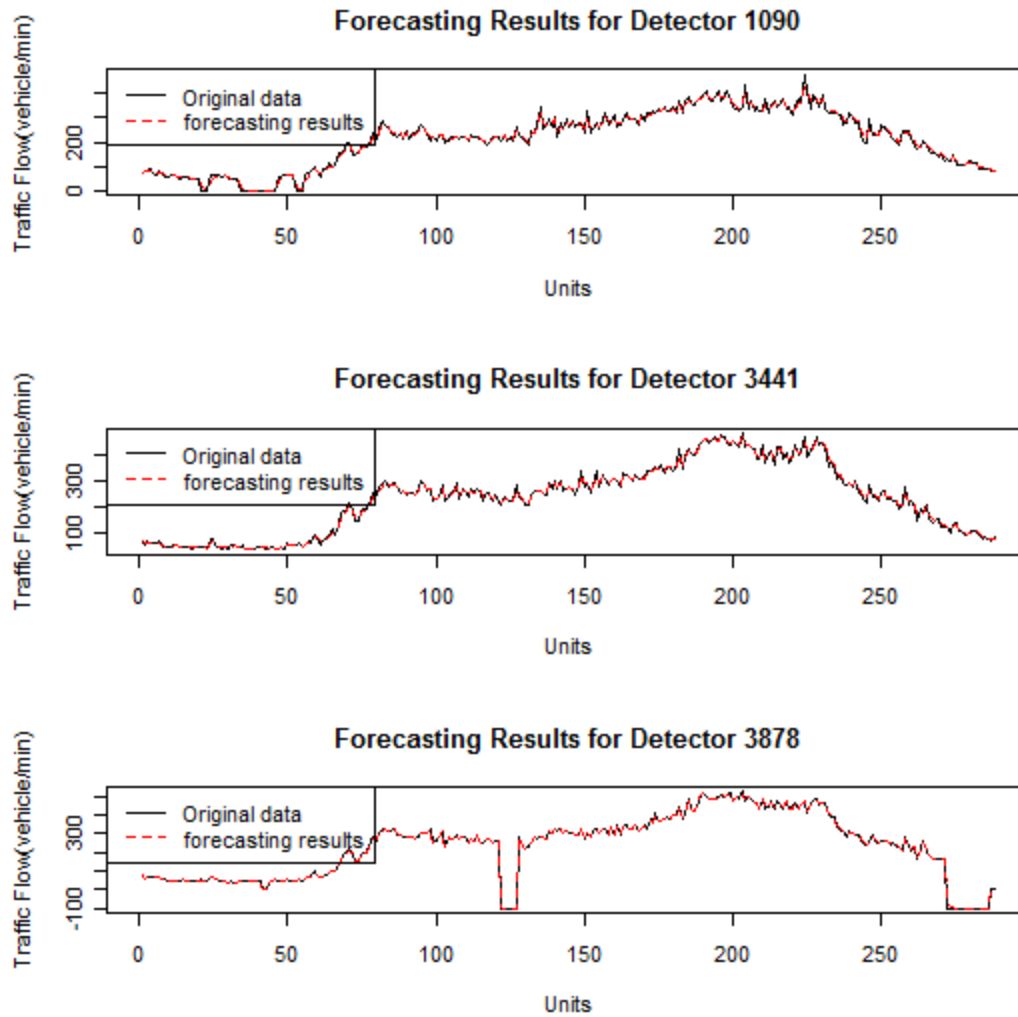
Figure 10. Forecasting results of ARIMA based traffic flow forecasting model(1)
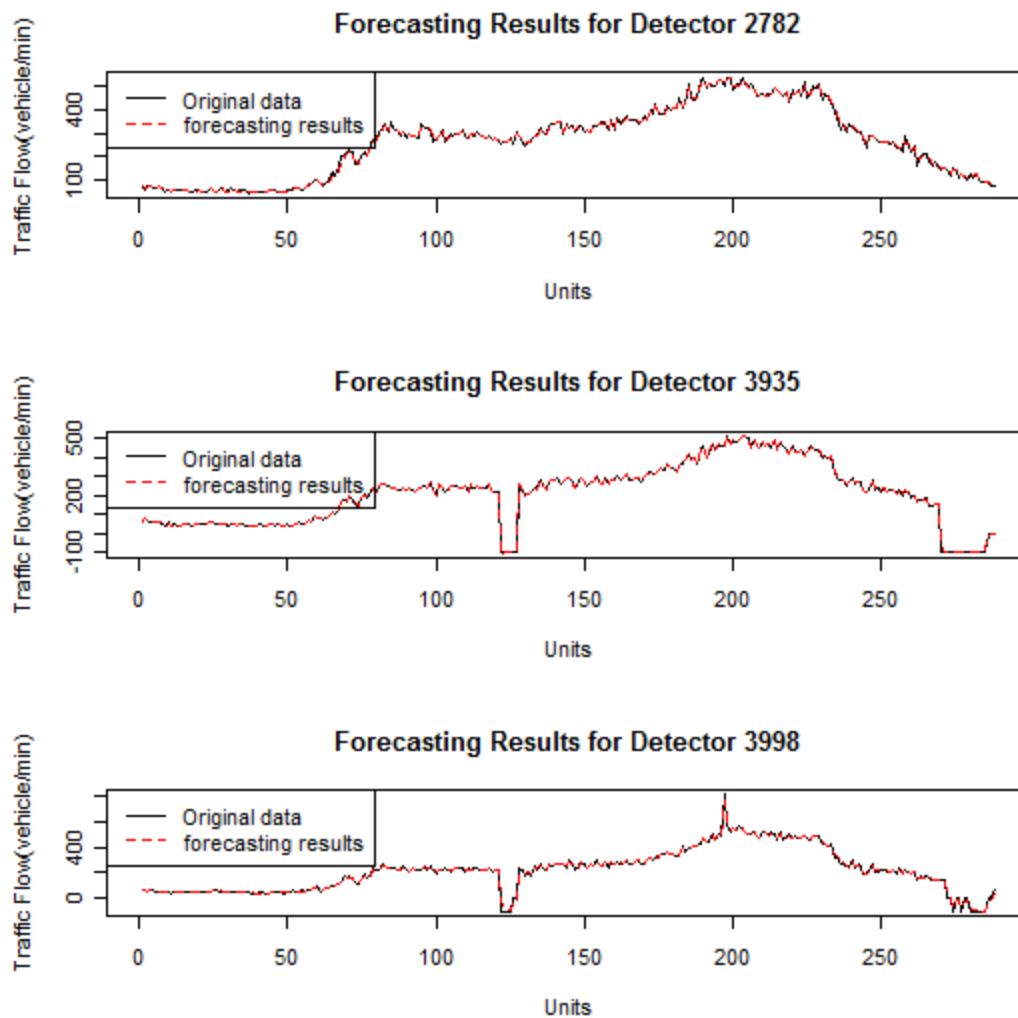
Figure 11. Forecasting results of ARIMA based traffic flow forecasting model(2)

Forecasting results of the ARIMA model for the six studied sites are represented in these figures, a red dash line represents the one step forecasting results and the black line is the field data obtained from detectors. These figures show that the ARIMA model can provide adequate forecasting results based on the historical traffic flow information. However, as we inspect forecasting results for detectors 3878, 3935 and 3998 carefully, most original traffic flow data points (the black line) were below zero during time steps

120 to 130 and from 270 to 288. The forecasted results at these points were also

approximately zero. This indicates that if missing data exits for a particular time periods,

the forecasting results will be affected. ARIMA model can give us very nice forecasting

results only if the historical data we obtained is complete and accurate.

5.1.2 GARCH Model Fitting

The previous section indicates that ARIMA model is capable at capturing the

expected values of traffic flow data. During non-peak traffic flow conditions, the

variance of the flow data is very small and the expected value of traffic flow data can be

approximately the actual value of traffic flow data. However, in accident, peak, and

abnormal traffic flow situations, the variance of traffic flow data can be very large. Only

relying on the expected forecasting value cannot provide adequate information either for

road users or traffic operation mangers to make proper decision. It is critical for us to

know whether if there is a big jump in traffic flow variation. By taking into

consideration other additional information (for example: speed data), one can figure out

the traffic conditions for the next time step. Thus, in this section, we focus on the

application of GARCH model.

The first step is to plot the residuals of the ARIMA model. The upper left plot in

Figure (12) indicates that the residuals are not white noise and certain patterns still exist

in the dataset. In order to further check if some patterns exist in the data, sample ACF

and PACF of various functions of residuals are plotted. The upper right figure is the

ACF plot of residuals series. It suggested that there are no serial correlations. The lower

left figure is the absolute value of the residuals while the lower right figure is the

squared value of the residuals. These two plots suggested that residuals are not serially

independent. All these three plots suggested that the residuals are serially uncorrelated

but dependent. To capture such dependency in residuals leads to more accurate

forecasting results.
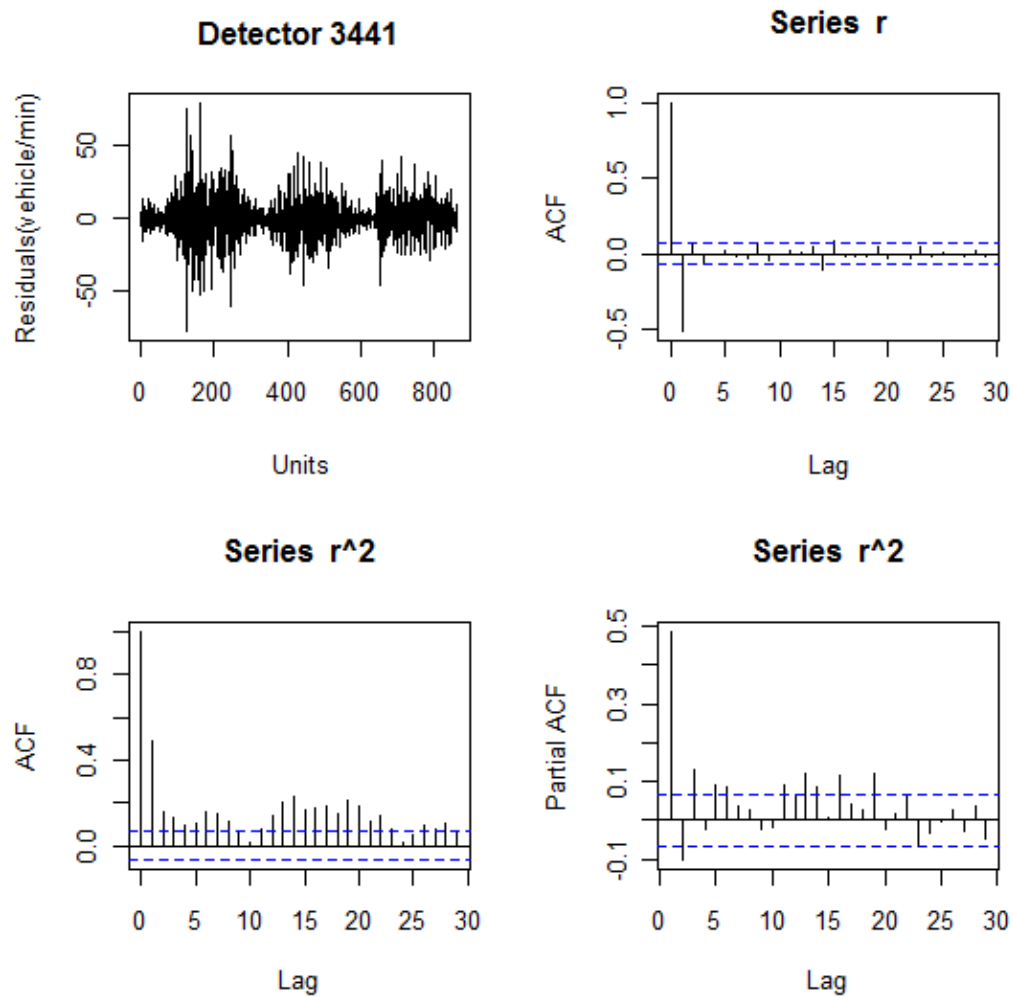
Figure 12. Residual analysis

A GARCH model is used to analyze the variance part of the traffic data. The

basic idea of the volatility model (GARCH model) is to find a mean structure model first

and then apply the GARCH model to the residual part. In this section, we will use

ARIMA (0, 1, 1) to represent the expected value of traffic flow data and apply the

GARCH model to predict the confidence interval of the forecasting results. One step

forecasting strategy is used.

If a ARIMA(0,1,1) model is used, the expected value of data is:

$$x_t = z_t + \theta z_{t-1} + x_{t-1} + e_t \tag{24}$$

For the variance part, we use the ARCH model:

$$e_t = z_t \sigma_t \tag{25}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^m \alpha_i e_{t-i}^2 + \sum_{j=1}^s \beta_j \sigma_{t-j}^2 \tag{26}$$

The equation of the joint the ARIMA(0,1,1)-GARCH(1,1) model then gives

$$x_t = z_t + \theta z_{t-1} + x_{t-1} + e_t \tag{27}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 e_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{28}$$

In the GARCH (1, 1) model, three parameters needed to be estimated: $\alpha_0, \alpha_1, \beta_1$.

The maximum likelihood parameter estimation method is used to choose the best

parameters. Figure (13) is the prediction confidence interval for the residual part of

traffic flow for Detector 3441. As we can see from this figure, residuals of traffic flow

continue to be large during certain time periods and the prediction interval has the ability

to capture the volatility nature of the data set. For example, there is a big change around

time step 60, the residuals drop below -20, then rise up and drop down again. The predicted confidence intervals show three peaks during this period which give us an indication that confidence band that contains the true value of the forecasted traffic flow during this period will be larger. The GARCH model provides direct information on how reliable the forecasting results are. If we take other information into consideration, such as speed or density, then we can figure out possible traffic conditions within the next five minutes.
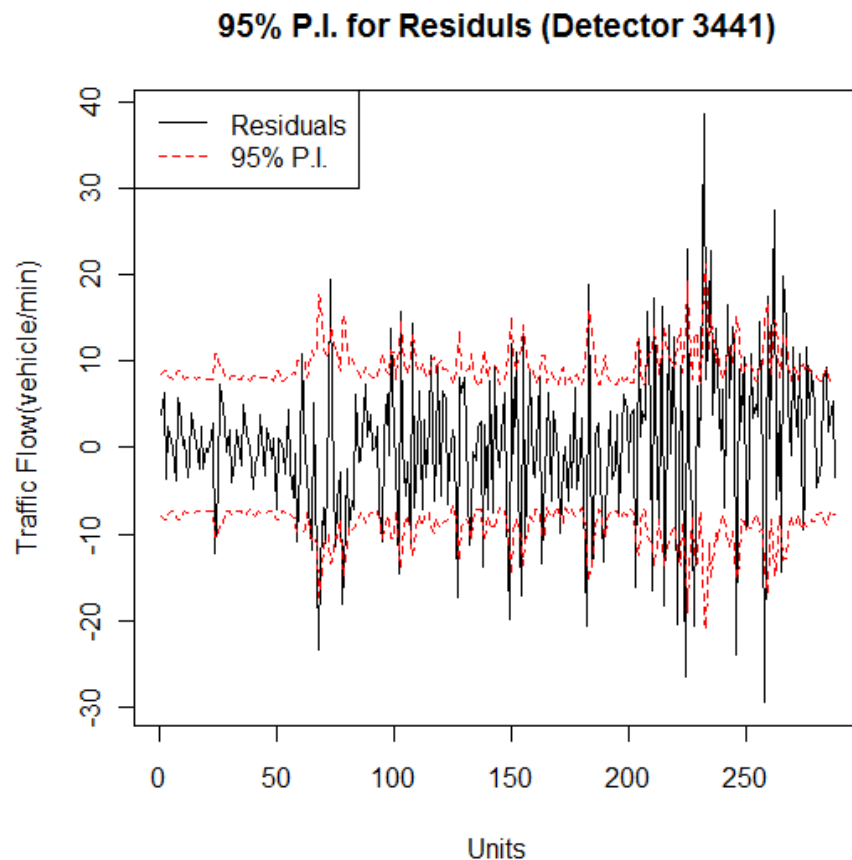
Figure 13. VAR forecasting results—95% prediction interval

5.2 Multivariate Forecasting

Although the univariate model provides promising forecasting results, it cannot deal with the missing data situation. As we have discussed before, when data is missing, the forecasting accuracy will be affected due to the fact that the univariate forecasting method only considers information from one detector. If data are not available for a certain time periods, next time step forecasting cannot be used. The commonly used methodology dealing with missing data is to take historical average. However, this method lacks theoretical support and the forecasting accuracy cannot be guaranteed. Considering the fact that special relationships exists among traffic flow data from different detectors, we use the multivariate model to deal with missing data situations.

Two methods are proposed for missing data situations: the VAR based method and the GRNN based method. The VAR based forecasting method uses traffic flow data from two detectors: the detector that has missing data and its up-stream counterpart, as the model input to forecast the next time step traffic flow. Thus, the forecasting result will be based upon traffic information from both its own time series data and the time series data from its up-stream counterpart. The VAR based method assumes a linear relationship between the two traffic flow series from the closely spaced detectors. The structure of the VAR model is simple and the forecasted value can be represented as a

linear combination of two time series. The GRNN based method forecast next time step traffic flow information for the studied site by only using up-stream traffic information as model input. The forecasting results are only based on its upstream information. Performance of these two models in data missing situations will be studied in the following sections.

5.2.1 Vector Autoregression (VAR) Model Fitting

Given the fact that the ARIMA model fits a univariate model well, an extension of the univariate autoregressive model-the vector autoregression (VAR) model will be the good choice for multivariate traffic flow forecasting. In this study, we will focus on the bivariate vector autoregressive model with two dependent time series: traffic flow at the up-stream and at the studied site. We divided the six studied sites into three groups: Detectors 1090 and 3441 as group one, detectors 3878 and 2782 as group two, detectors 3935 and 3998 as group three. Then maximum-likelihood estimation (MLE) method is used for parameter estimation. Then a one step ahead forecasting strategy is used to predict traffic flow on each group. Figure 14 to Figure 16 show the forecasting results:
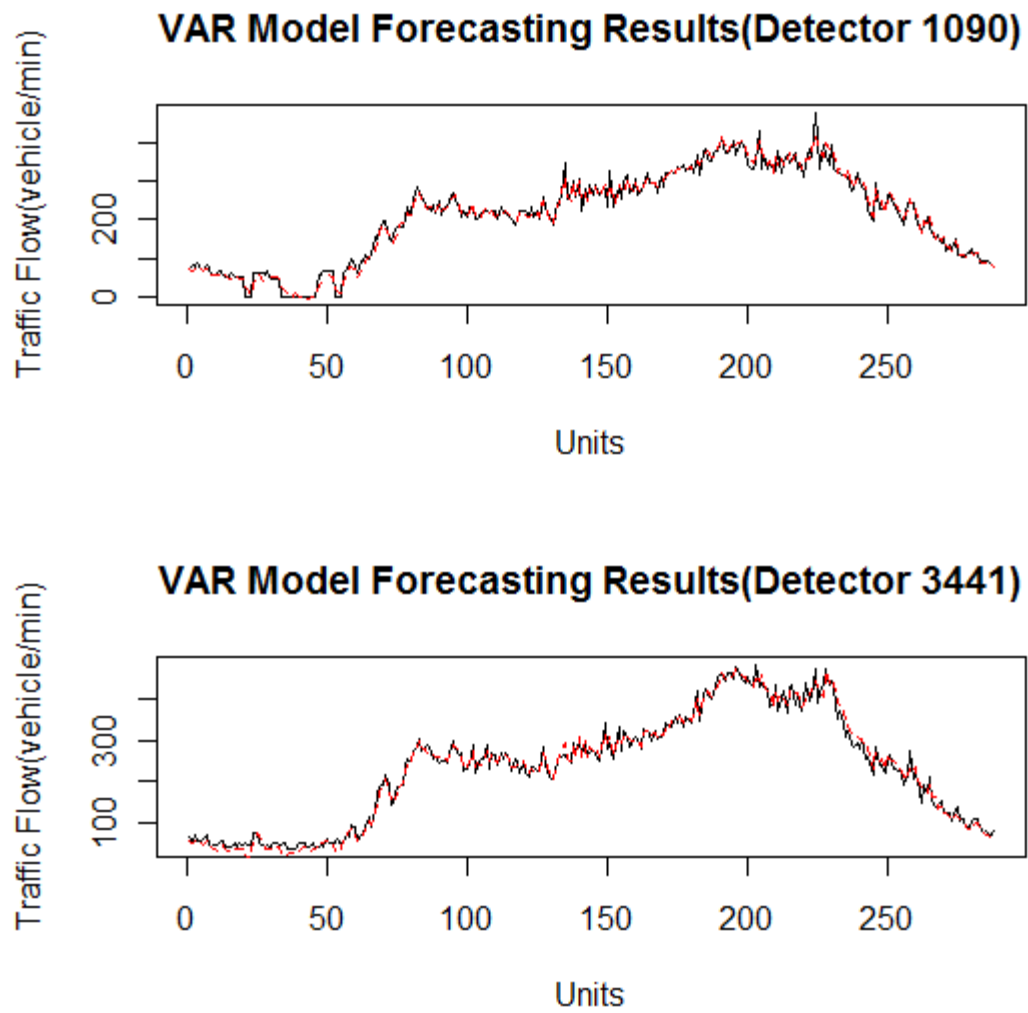
Figure 14. VAR model forecasting results(1)
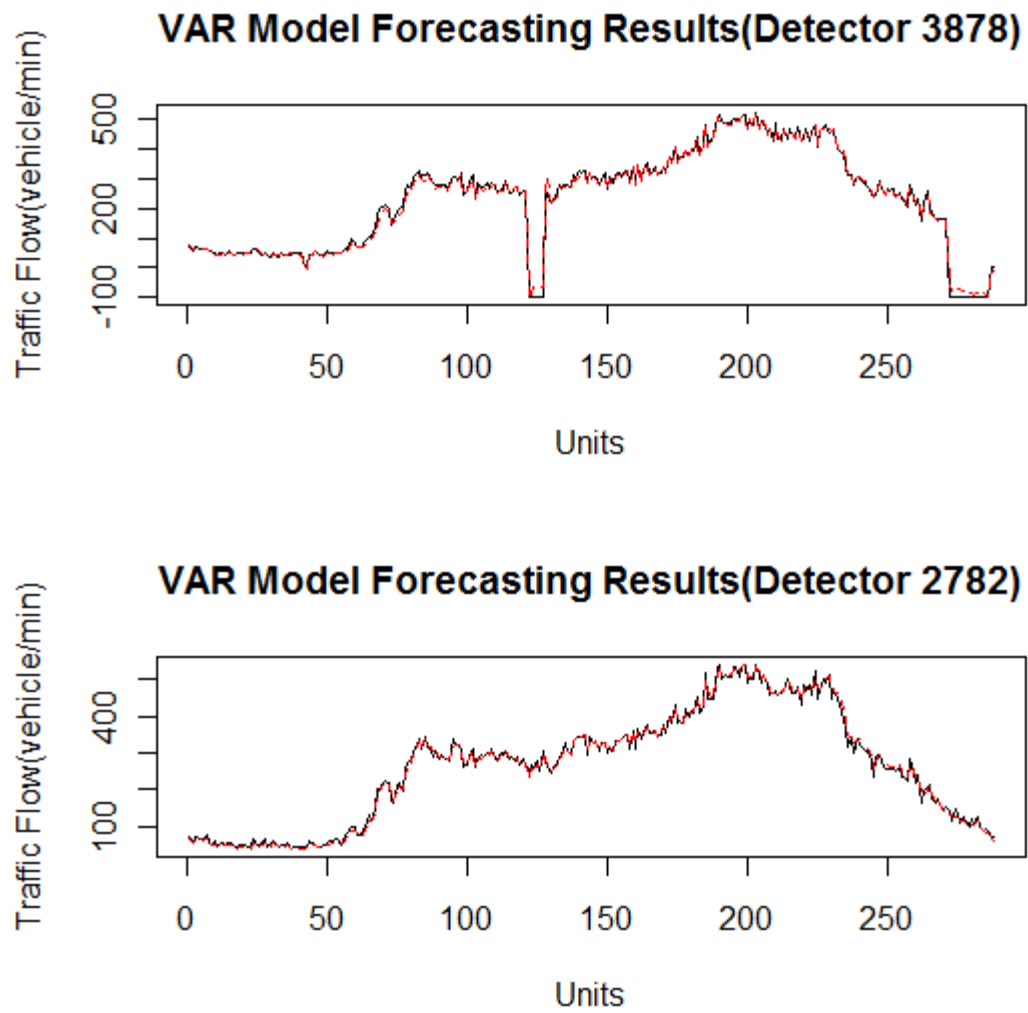
Figure 15. VAR model forecasting results(2)
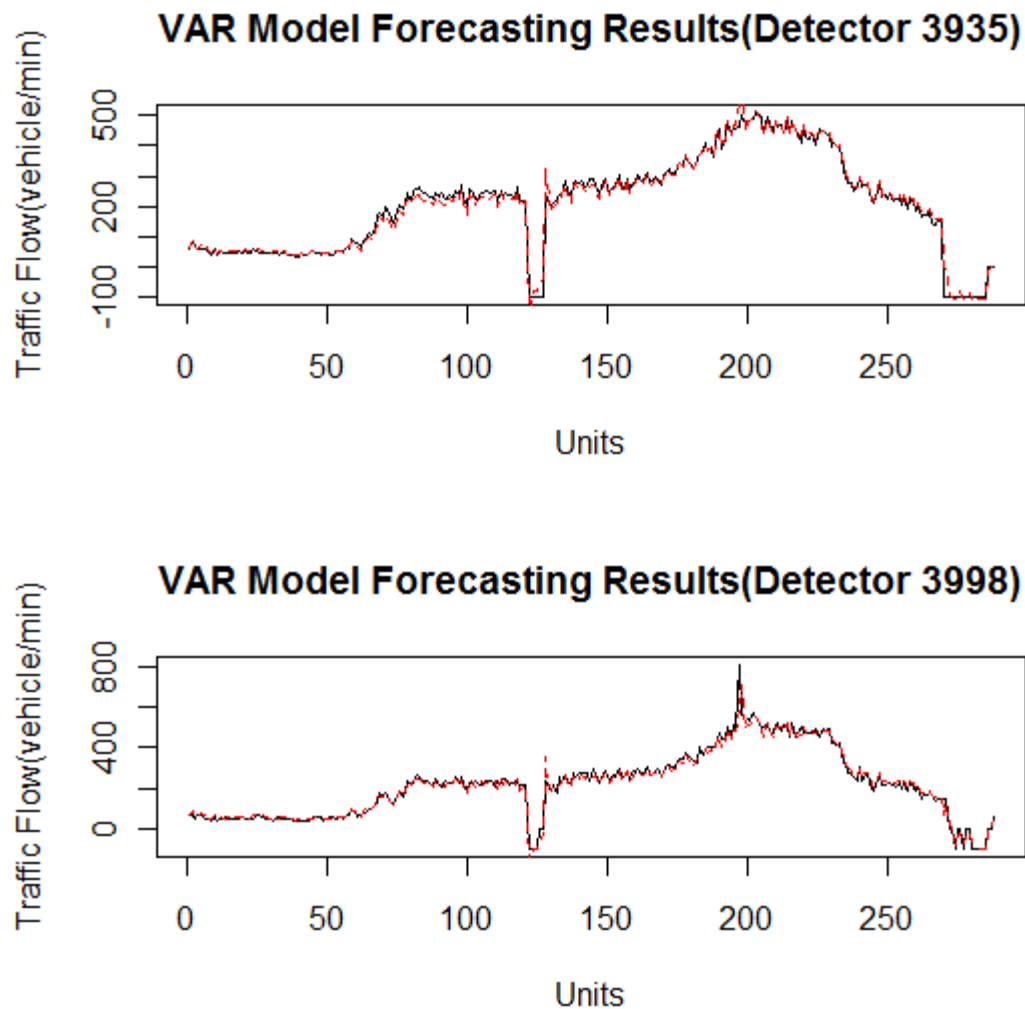
Figure 16. VAR model forecasting results(3)

These three plots represent three different situations: no missing data exists for both time series (group one), one detector has missing data (group two) and both detectors have missing data (group three). A cursory look of these three plots indicate that: VAR model can provide adequate forecasting of traffic flow data in the next time step when no missing data exits; If only one studied series has missing data, the

forecasted value will be influenced by the other time series and thus will be higher than zero; If both time series have missing data for the same interval, the forecasting result during data missing period will stay around or below zero.

Although the vector autoregressive model takes into consideration the flow information from other detectors, the forecasting results are still being affected by the missing data. Since the forecasted value is a linear regression function of its own past values and the past values of another time series, it will drop down if one or two variables go below zero. Another disadvantage of the VAR model is that it can only represent the linear relationship among different variables. However, if a nonlinear relationship exists between two traffic flow series, it is important to take this into consideration when conduct traffic flow forecasting.

5.2.2 General Regression Neural Network (GRNN) Model Fitting

The GRNN model belongs to the category of probabilistic neural networks, which only need a smaller fraction of the training samples compared to back propagation neural networks. The advantage of the GRNN model is that it converges to the underlying function of the data without preliminary knowledge of the data. It is a very useful tool to perform predictions.

As we mentioned before, spatial correlations exist among traffic flow data. Up-stream traffic information can be used to predict down-stream traffic information in the next few time steps. If there is missing data for the studied site, traffic information from its up-stream location can be used to perform forecasting without information at the studied site. In this section, we will focus on using up-stream traffic flow information to

predict traffic at the studied site to deal with the missing data problem. The first step is model training. In this step, historical traffic flow data from both upstream detector and detector of interest are used as model input and model output respectively. This step is aimed at training the neural network. The second step is forecasting. Once the model training is completed, one can use the current upstream traffic flow information as model input to predict future traffic flow information at a studied detector. Figure 17 shows the GRNN model development process.
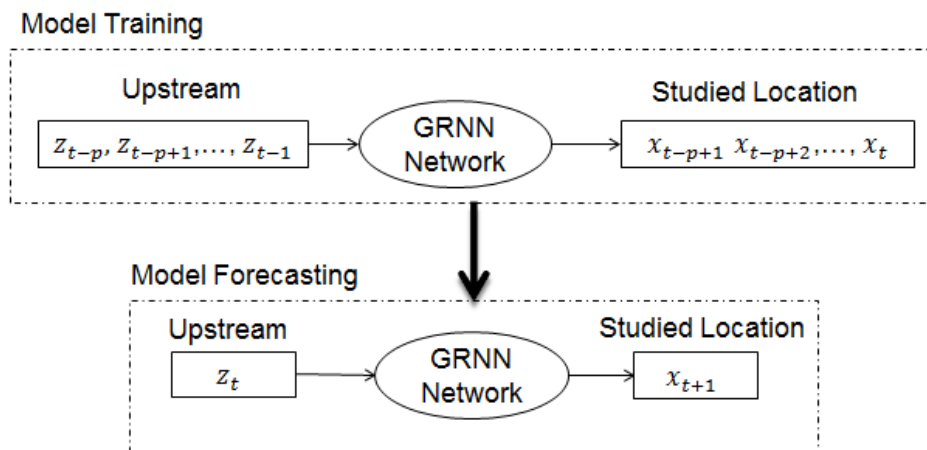
Figure 17. GRNN model development

Detectors 3878, 3935 and 3998 are chosen as the locations of interest, since they all have missing data on February 5th, 2008. The forecasting strategy is that if there is missing data for the studied detector, then one should search historical traffic flow data for its up-stream detectors. If missing data from the nearest detector is not significant, then traffic flow data at this detector will be used as the input of the GRNN model and historical traffic flow data at the studied detector will be the output of the forecasting model. If missing data is also a problem for its nearest up-stream detector, the procedure is to find another nearest detector that does not have missing data. In this study, traffic flow data from Detector 3441 will be used to predict traffic information for Detector 3878, traffic flow information from Detector 2782 will be used to forecast traffic flow at Detector 3935, and Detector 2782 will be used to predict flow at Detector 3998. Figure 18 is the forecasting results for Detectors 3878, 3935, and 3998.It indicates that the predicted values fit the original data well. Unlike the ARIMA and VAR based forecasting methods, the GRNN forecasting results are not affected by the missing data since it is only based on the history data from its up-stream detector.
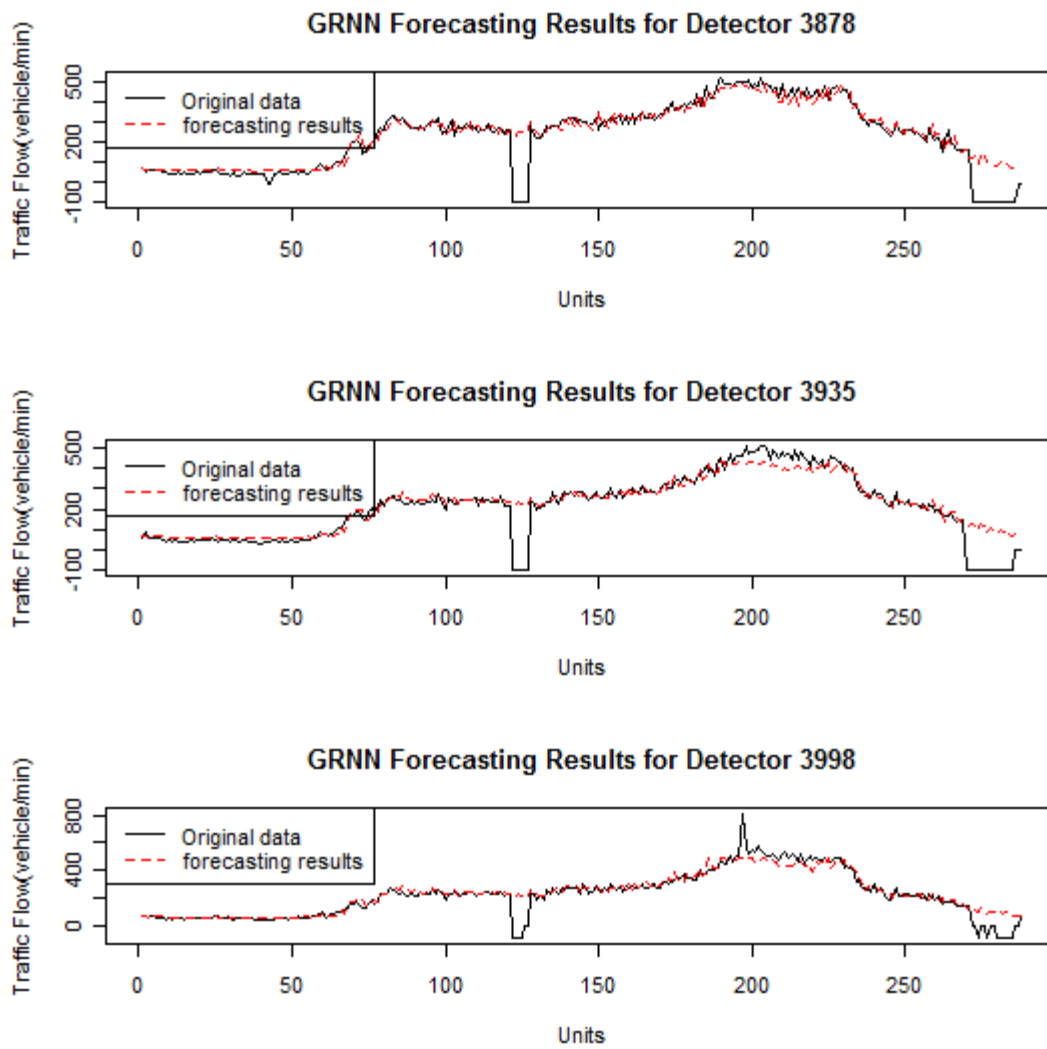
Figure 18. GRNN forecating results for three detectors that have missing data

5.2.3 Historical Average Model Fitting

The historical average model simply uses the average value of historical traffic flow data to represent future traffic volume. It is based on the seasonal characteristic of traffic flow data, e.g. traffic flow patterns day to day often show remarkable similarity and these patterns are useful for prediction. The historical average method is easy to

understand and implement. It has already been applied to the urban traffic control systems (UTCS) (Stephanedes et al. 1981) and other various traveler information systems (Jeffrey et al. 1987 and Kaysi et al. 1993). However, it only relies on past traffic information and cannot react to dynamic changes of traffic flow.

The presented model in this study is to find out the average value of past traffic volume for each time interval and each site. For example, if we want to predict traffic flow at time t on February 5, 2008 we take average of traffic volume at time t in previous days. In this study, 35 day traffic volume information is used to predict traffic flow on February 5, 2008 and the results of this method will be discussed in section 6.

6.   MODEL COMPARISON AND ANALYSIS

This research studies the ARIMA-GARCH, the VAR, the GRNN, and the Historical Average models. Moreover, this study also addresses traffic flow forecasting reliability and missing data. The ARIMA-GARCH model is aimed at improving forecasting accuracy and reliability in non-missing data situations. In addition, he VAR, the GRNN, and the Historical Average models are applied in dealing with missing data situations.

First, model comparison in non-missing data situations are studied. In this part, forecasting accuracies of four proposed models are studied. Then, the study also discusses model performance in missing data situations. This section presents strengths and weaknesses of each model and discusses how to choose a proper model in a certain situation.

6.1 Model Comparison in Non-Missing Data Situations

In order to compare the forecasting accuracy in normal conditions (no missing data) numerically, there are two measures of effectiveness: the root mean squared error (RMSE) and the mean absolute percentage error (MAPE). The RMSE is representative of the size of a "typical" error because it is measured in the same unite as the original data. It is more common than the mean squared error (MSE). The MAPE is another commonly used measure of effectiveness for purposes of reporting because it is

expressed in percentage terms, which give us a general sense of the error even without

knowledge of what constitutes a "big" error for the data set.

The equation of RMSE is:

$$RMSE = (\sum_{i=1}^{n} \frac{(t(i) - a(i))^2}{n})^{1/2} \quad, \qquad (29)$$

while the equation of MAPE is

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{t(i) - a(i)}{t(i)} \right| \quad, \qquad (30)$$

where $t(i)$ is the actual value, $a(i)$ is the forecast value ,and n is the total number of data

intervals.

Table 2. and Table 3. investigate the RMSE and the MAPE values for four

models: the ARIMA-GARCH, the VAR, the GRNN, and the Historical Average models.

Because one cannot obtain the true values of traffic flow data at missing data points, we

omit missing data before calculating the RMSE and MAPE value. From Table 2., the

RMSEs for each detector from the ARIMA-GARCH model are better than the other

three models. Table 3.also indicates that the ARIMA-GARCH model outperforms the

other three models based on the MAPEs criterion. From these results, one can conclude

that ARIMA-GARCH model performs best among these three models in non-missing

data situations. As we already discussed in Section 5, the GARCH model is capable of

modeling variance part of traffic flow. Thus, the GARCH model can provide

information on how reliable the forecasting accuracy is. The ARIMA-GARCH model

provides the best forecasting results among the studied models and it performs well in

non- missing data situations in traffic flow forecasting.

Table 2 RMSE values of four forecasting methods

RMSE						
	D1090	D3441	D3878	D2782	D3935	D3998
ARIMA-GARCH	11.78	9.53	4.12	7.19	2.93	7.92
VAR	15.99	13.36	7.66	9.55	12.06	22.37
GRNN	--	--	27.82	--	28.69	34.35
HA	--	--	59.80	--	42.17	65.24

Table 3 MAPE values of four forecasting methods

MAPE						
	D1090	D3441	D3878	D2782	D3935	D3998
ARIMA-GARCH	4.61%	3.45%	1.24%	2.33%	0.92%	2.30%
VAR	6.61%	6.00%	2.73%	3.61%	6.06%	6.08%
GRNN	--	--	11.93%	--	13.65%	10.48%
HA	--	--	16.65%	--	15.56%	16.52%

65
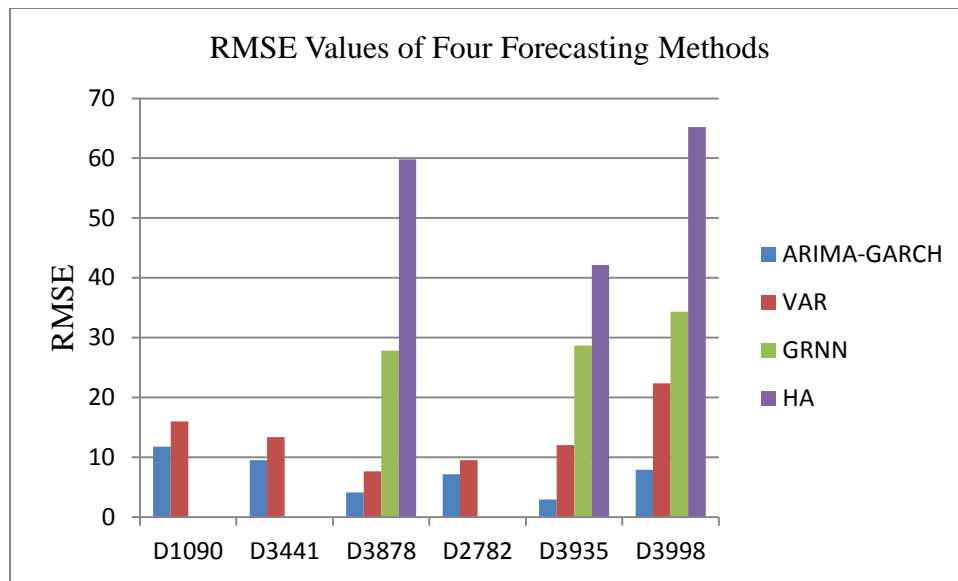
Figure 19. RMSE values plots of four forecasting methods
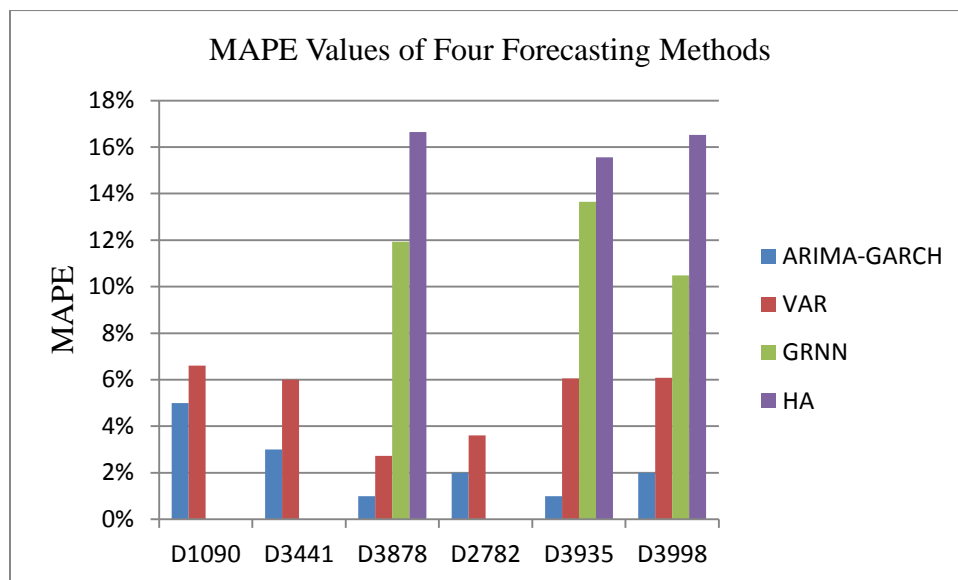
Figure 20. MAPE values plots of four forecasting methods

6.2 Discussion of Model Performance in Missing Data Situations

Although the ARIMA-GARCH model provides the best forecasting results

among the four proposed models, missing data will affect its performance. This can be

referred in Figure 10. and Figure 11., the numbers -1 or -99 represent the missing data. Due to the factor that the ARIMA-GARCH model only relies on historical and current flow information on its own time series, it does not have the ability to deal with missing data situations. If there is missing data for a certain point, the ARIMA-GARCH model will forecast the future traffic flow based on the values at the missing data points (-1 or -99). As indicated in Figure 10. and Figure 11., forecasted traffic flow values go below zero when there is missing data and this situation, which cannot happen in real life.

This paper proposes the VAR, the GRNN, and the Historical Average models to deal with missing data situations. The VAR model forecasts future traffic flow by considering historical and current traffic flow information from both its own data sets and the data from its up-stream location. Figure 14., Figure 15. and Figure 16. are the forecasting results of the VAR model. Although the VAR model has the ability to represent a linear relationship among traffic flow information at different locations, it does not perform well in missing data situations. In these figures, some forecasted values go below zero when there is missing data. Unlike the ARIMA-GARCH and the VAR models, the GRNN based forecasting method studies the traffic flow relationship between two detectors and forecast future traffic flow by only using flow information from the up-stream detector. The GRNN model did better in missing data situations as one can see from Figure 18. Although there is missing data in historical and current traffic flow data at the studied site, the GRNN forecasting results are based on traffic flow information from its upstream location and are not affected by the missing data. This model has the potential for dealing with missing data situations. In literature,

another commonly used method in dealing with missing data situations is the Historical Average method. In this study, an average of 35 days historical data(from January 1 to February 4) at each time point are taken to predict traffic flow information at its corresponding time point in February 5. Because both the Historical Average method and the GRNN method do not rely on traffic flow information from February 5, we can assume no flow information from February 5 is available when forecasting is performed. As indicated in Table (2) and Table (3), the GRNN model outperforms the Historical Average model based on RMSE and MAPE criteria. Thus, the GRNN model has the potential to deal with missing data situations.

7. CONCLUSIONS

This study addresses Traffic flow forecasting accuracy and missing data problems. First, this study introduces the volatility model to study the variance part of the traffic flow data, because it has the ability to indicate whether or not there is a big change in traffic flow over the next few minutes. By providing prediction confidence band for future traffic flow, one can capture the uncertainty of traffic flow forecasting results. Second, this study uses the Multivariate methods to ease the missing data problem. Two multivariate methods are proposed: the Autoregressive Vector model and the General Regression Neural Network model, to forecasting traffic flow in both normal and data missing situation. The following part summarizes the findings and conclusions of this research:

1. Seasonal component exists in traffic flow data, which can be removed by one-step difference of the original data. The differenced traffic flow data are one-step correlated. In other word, the increase or decrease of traffic flow data can influence the change of traffic flow data in the next time step. The prediction of traffic flow data can be made simpler by studying the differenced original traffic data.

2. The ARIMA-GARCH model fits the historical traffic flow data well and outperformed the VAR and GRNN models in non-missing data situations. However, there is missing data in historical traffic flow data and it will affect the

forecasting accuracy. Since the idea of the ARIMA-GARCH model is that forecasting of future traffic flow is based on the historical traffic flow data, and in the ARIMA-GARCH model, the next step forecasting results is closely related with its current traffic data. As a result, missing data will influence the forecasting accuracy.

3. The VAR model, an extension of the univariate autoregressive model, uses multiple traffic flow data to forecast and can represent the correlations of multiple times series. It is very useful if knowledge of correlation among multiple times series is needed. However, if one only considers forecasting accuracy, the VAR model did not perform well compared with the ARIMA model in this study.

4. The GRNN model has the ability to use upstream traffic information to predict studied site. Although the forecasting accuracy is not as good as the ARIMA-GARCH model in normal situations, it outperforms the ARIMA model and the VAR model in missing data situations. Because future traffic flow forecasting results of the studied site are solely based on its upstream traffic information, even if data are missing at the studied site, the forecasting results will not be affected.

8. LIMITATIONS AND FUTURE WORK

In this study, traffic flow forecasting reliability in non-missing data situations and traffic flow forecasting accuracy in missing data situations are studied. The study results indicate that the ARIMA-GARCH model outperforms other methods in non-missing data situations, while the GRNN model performs better in missing data situations. Since both models have their own advantages in different situations, the future work is to combine these two models to deal with both missing data and non-missing data situations.

REFERENCES

Ahmed, M. S. and Cook, A. R. (1979). "Analysis of freeway traffic time-series data by using Box-Jenkins technique." Transportation Research Record, 722, 1-9.

Bollerslev, T. (1986). "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, 31, 307-327.

Box, G.E.P. and Jenkins, G. M. (1970). "Time series analysis: Forecasting and control." *Holden-Day*, San Francisco.

Castro-Netoa, M.,Jeongb, Y.S., and Jeongb,M.K. , and Hana,L.D. (2009). "Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions." *Expert Systems with Applications*, 36(3), 6164-6173.

Chang, S. C., Kim, R. S., Kim, S. J., and Ahn, B. H. (2000). "Traffic-flow forecasting using a 3-stage model." *IEEE Intelligent Vehicle Symposium*, 451-456.

Chen, L. and Wang, F.Y. (2002). "A neuro-fuzzy system approach for forecasting short-term freeway traffic flows." *IEEE 5th International Conference on Intelligent Transportation Systems*, Singapore: IEEE Intelligent Transportation Systems Council, 747-751.

Chen, S.Y., and Wang, W. (2006). "Traffic volume forecasting based on wavelet transform and neural networks." *3rd International Symposium on Neural Networks*. Chengdu, China, 3973, 1-7.

Davis, G.A. and Nihan, N.L. (1984). "Using time-series designs to estimate changes in freeway level of service, despite missing data." *Transportation Research Part A: General*, 18(5), 431-438.

Dougherty, M. (1995). "A review of neural networks applied to transport." *Transportation Research Part C: Emerging Technologies*, 3(4), 247-260.

Guin, A. (2004). "An incident detection algorithm based on a discrete state propagation model of traffic flow." PhD Dissertation, Georgia Institute of Technology, Atlanta, GA.

Guo, J., Williams, B. M., and Smith, B. L. (2007). " Data collection time intervals for stochastic short-term traffic flow forecasting." *Transportation Research Record: Journal of the Transportation Research Board*, 2024, 18-26.

"Highway capacity manual." 2000. Transportation Research Board, Washington, DC.

Jeffrey,D.J.,Russam,K., and Robertson,D.I. (1987). "Electronic route guidance by AUTOGUIDE: The research background." *Traffic Engineering and Control*, 525-529.

Jin,F. and Sun,S. (2008). "Neural network multitask learning for traffic flow forecasting." *IEEE World Congress on Computational Intelligence*, 1897-1901.

Kamarianakis, Y., and Prastacos, P. (2003). "Forecasting traffic flow conditions in an urban network: Comparison of univariate and multivariate approaches." *Transportation Research Record: Journal of the Transportation Research Board*, 1857, 74-84.

Kamarianakis, Y., Kanas, A., and Prastacos, P. (2005). "Modeling traffic volatility

dynamics in an urban network." *Transportation Research Record: Journal of the*

Transportation Research Board, 1923, 18-27.

Kaysi,I.,Ben-Akiva,M., and Koutsopoulos,H. (1993). "An integrated approach to vehicle

routing and congestion prediction for real-time drvier guidance." *Transportation*

Research Record,1408, 66-74.

Lam, S.H.M. and Toan,T.D. (2008). "Short-term travel time prediction using support

vector regression." *Transportation Research Record: Journal of Transportation*

Research Board, 08-0670,16.

Ledoux, C. (1997). "An urban traffic flow model integrating neural networks."

Transportation Research Part C: Emerging Technologies, 5(5), 287-300.

Lieu,H. (1999) " Revised monograph on traffic flow theory."

<http://www.fhwa.dot.gov/publications/research/operations/tft/index.cfm> (April

15,2011)

Nau,B. (2005). "Introduction to ARIMA: Nonseasonal models." Decision 411

Forecasting. <http://www.duke.edu/~rnau/411arim.htm> (April 5,2011).

Newell, G.F. (1993). "A simplified theory of kinematic waves in highway traffic, part I:

General theory." *Transportation Research Part B: Methodological*, 27(4), 281-287.

Nicholson, H. and Swann, C. D. (1974). "The prediction of traffic flow volumes based

on spectral analysis." *Transportation Research*, 8(6), 533-538.

Okutani,I., and Stephanedes,Y.J. (1984). "Dynamic prediction of traffic volume through Kalman filtering theory." *Transportation Research Part B: Methodological*,18(1), 1-11.

Park, B. (2002). "Hybrid neuro-fuzzy application in short-term freeway traffic volume forecasting." *Transportation Research Record*, 1802,190-196.

Park, B., Messer, C. J., and Thomas, U. II. (1998). "Short-term freeway traffic volume forecasting using radial basis function neural network." *Transportation Research Record: Journal of the Transportation Research Board*, 1651, 39-41.

Pfeifer, P.E., and Deutsch, S.J. (1980). "A three-stage iterative procedure for space–time modeling." *Technometrics*, 22(1), 35-47.

Said , S. E. and Dickey, D. A. (1984). "Testing for unit roots in autoregressive-moving average models of unknown order." *Biometrika,* 71, 599-607.

Smith, B. L., and Demetsky M. J. (1994). "Short-term traffic flow prediction: Neural network approach." *Transportation Research Record: Journal of the Transportation Research Board*, 1453, 98-104.

Sohn, K., and Kim, D. (2009). "Statistical model for forecasting link travel time variability." *ASCE Journal of Transportation Engineering*, 135(7), 440-453.

Stephanedes, Y.J., Michalopoulos,P.G., and Plum,R.A. (1981). "Improved estimation of traffic flow for real-time control." *Transportation Research Record*, 795, 28-39.

Sun, S. and Zhang, C. (2007). "The selective random subspace predictor for traffic flow forecasting." *IEEE Transactions on intelligent transportation systems*, 8(2), 367-373.

Szeto, W.Y., Ghosh,B., Basu,B., and Mahony,M.O. (2009). "Multivariate traffic forecasting technique using cell transmission model and SARIMA model." *J. Transp. Engrg.,* 135, 658.

Tsekeris, T. and Stathopoulos, A. (2006). "Real-time traffic volatility forecasting in urban arterial networks." *Transportation Research Record: Journal of the Transportation Research Board*, 1964, 146-156.

Voort,M.D., Dougherty,M., and Watson,S. (1996). "Combining Kohonen maps with ARIMA time series models to forecast traffic flow." *Transportation Research Part C,* 4(5), 307-318.

Wild, D. (1997). "Short-term forecasting based on a transformation and classification of traffic volume time series." *International Journal of Forecasting,* 13(1), 63-72.

William, B.M. (1999). "Modeling and forecasting vehicular traffic flow as a seasonal stochastic time series process." Doctoral Dissertation, University of Virginia, Charlottesville.

Williams, B. M., Durvasula, P. K., and Brown, D. E. (1998). "Urban freeway traffic flow prediction: Application of seasonal autoregressive integrated moving average and exponential smoothing models." *Transportation Research Record*, 1644, 132-141.

Xiao, H., Sun, H., Ran, B., and Oh, Y. (2003). "Fuzzy-neural network traffic prediction framework with wavelet decomposition." *Transportation Research Record*, 1836, 16-20.

Ye, Z.R., Zhang, Y.L., and Middleton, D. R. (2006). "Unscented Kalman filter method for speed estimation using single loop detector data." *Transportation Research Record*, 1968, 117-125.

Yin, H. B., Wong, S. C., Xu, J. M., and Wong, C. K. (2002). "Urban traffic flow prediction using a fuzzy-neural approach." *Transportation Research Part C: Emerging Technologies*, 10(2), 85-98.

VITA

Name: Yanru Zhang

Address: c/o Department of Civil Engineering

                Texas A&M University, 3136

                College Station, TX 77840

Email Address: zhangyanru_anne@tamu.edu.

Education: B.S., Transportation Engineering, Beijing Jiaotong University, 2009

                M.S., Transportation Engineering, Texas A&M University, 2011