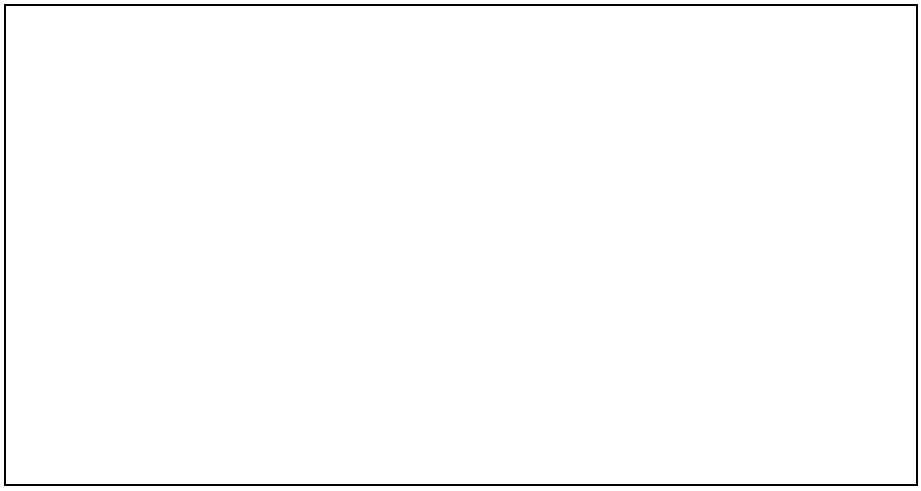


The Pennsylvania State University ❖ University of Maryland
University of Virginia ❖ Virginia Polytechnic Institute and State University
West Virginia University

The Pennsylvania State University ❖ The Thomas D. Larson Pennsylvania Transportation Institute
Transportation Research Building ❖ University Park, Pennsylvania 16802-4710
Phone: 814-863-1909 ❖ Fax: 814-863-3707
www.pti.psu.edu/mautc

ESTIMATION AND PREDICTION OF ORIGIN-DESTINATION MATRICES FOR I-66

FINAL REPORT

VT-2009-02
DTRT07-G-0003

Prepared for

Mid-Atlantic University Transportation Center

By

Weihao Yin, Graduate Research Assistant
Pamela Murray-Tuite, Assistant Professor

Department of Civil and Environmental Engineering
Virginia Polytechnic and State University
Northern Virginia Center
Falls Church, VA, 22043

August 1, 2011

1. Report No. VT-2009-02	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Estimation and Prediction of Origin-Destination Matrices for I-66		5. Report Date September 2011
		6. Performing Organization Code
7. Author(s) Weihao Yin and Pamela Murray-Tuite		8. Performing Organization Report No.
9. Performing Organization Name and Address Department of Civil and Environmental Engineering Virginia Polytechnic and State University Northern Virginia Center Falls Church, VA, 22043		10. Work Unit No. (TRAIS) 11. Contract or Grant No. DTRT07-G-0003
12. Sponsoring Agency Name and Address US DOT Research & Innovative Technology Admin UTC Program, RDT-30 1200 New Jersey Ave., SE Washington, DC 20590		13. Type of Report and Period Covered Final Report 8/10/2008– 8/10/2011 14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

This project uses the Box-Jenkins time-series technique to model and forecast the traffic flows and then uses the flow forecasts to predict the origin-destination matrices. First, a detailed analysis was conducted to investigate the best data correction method.

Four spatial correction procedures were examined for non-incident related detector data. The first approach, temporal correction, exploited the inherent temporal trend of historical traffic. The spatial correction based on linear regression (LR) - a proposed modification of a previous approach - uses the relationship between the individual detector flow and station flow. The third approach proposed in this study is also a spatial correction method. A unique feature of the proposed spatial correction procedure was incorporation of lane use percentage into the correction process through kernel regression (KR). As a comparison benchmark, the correction method based on lane distribution (LD) developed by previous researchers was included as the fourth method.

To comprehensively compare the correction procedures, both systematic evaluation and random-error evaluation were conducted. After the results of systematic evaluation were analyzed, it was found that adaption was needed for the KR and LD approaches. Specifically, the individual lane flows provided by the detectors on particular general purpose lanes produced more accurate estimates. The two correction procedures (kernel regression and lane distribution) were revised in light of this finding and their station flow estimates were compared to those of the temporal correction and the LR approach at five error levels, which was considered as the random-error evaluation.

After applying the temporal correction method to the data set, the station flow series was modeled using the Box-Jenkins time-series modeling framework. The station flow series was successfully modeled by an MA(2) model using the Box-Jenkins time series technique. It enjoys an MAE of 344.53 and MAPE of 7.07%. The limitation is that different models are needed for each station and the model may be different for a different time period. However, once the model is fitted, it can be used to forecast the traffic hourly

flows, which were subsequently used to predict the origin-destination demands.

The performance of QueensOD modeling is relatively good, which is evidenced by approximately 10% MAPE for a majority of links. Generally, the relative errors (MAPE) are around 10% with a few higher exceptions. This is possibly due to the fact that the QueensOD requires the actual travel time as inputs. The actual travel times were computed using the BPR function according to observed flows, which could be a source of error.

<table>
<tr><td colspan="2">17. Key Words
Data Correction, OD Estimation, Time Series</td><td colspan="2">18. Distribution Statement
No restrictions.</td></tr>
<tr><td>19. Security Classif. (of this report) Unclassified</td><td>20. Security Classif. (of this page)
Unclassified</td><td>21. No. of Pages</td><td>22. Price</td></tr>
</table>

ABSTRACT

This project uses the Box-Jenkins time-series technique to model and forecast the traffic flows and then uses the flow forecasts to predict the origin-destination matrices. First, a detailed analysis was conducted to investigate the best data correction method.

Four spatial correction procedures were examined for non-incident related detector data. The first approach, temporal correction, exploited the inherent temporal trend of historical traffic. The spatial correction based on linear regression (LR) - a proposed modification of a previous approach - uses the relationship between the individual detector flow and station flow. The third approach proposed in this study is also a spatial correction method. A unique feature of the proposed spatial correction procedure was incorporation of lane use percentage into the correction process through kernel regression (KR). As a comparison benchmark, the correction method based on lane distribution (LD) developed by previous researchers was included as the fourth method.

To comprehensively compare the correction procedures, both systematic evaluation and random-error evaluation were conducted. After the results of systematic evaluation were analyzed, it was found that adaption was needed for the KR and LD approaches. Specifically, the individual lane flows provided by the detectors on particular general purpose lanes produced more accurate estimates. The two correction procedures (kernel regression and lane distribution) were revised in light of this finding and their station flow estimates were compared to those of the temporal correction and the LR approach at five error levels, which was considered as the random-error evaluation.

After applying the temporal correction method to the data set, the station flow series was modeled using the Box-Jenkins time-series modeling framework. The station flow series was successfully modeled by an MA(2) model using the Box-Jenkins time series technique. It enjoys an MAE of 344.53 and MAPE of 7.07%. The limitation is that different models are needed for each station and the model may be different for a different time period. However, once the model is fitted, it can be used to forecast the traffic hourly flows, which were subsequently used to predict the origin-destination demands.

The performance of QueensOD modeling is relatively good, which is evidenced by approximately 10% MAPE for a majority of links. Generally, the relative errors (MAPE) are around 10% with a few higher exceptions. This is possibly due to the fact that the QueensOD requires the actual travel time as inputs. The actual travel times were computed using the BPR function according to observed flows, which could be a source of error.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

This chapter outlines the background and motivations for this project. In addition, the problem statement is also provided to define the boundary of the research effort and highlight the tasks that constitute this research initiative.

1.1. Background and Motivations

Demand for freeway travel increases as the population grows and economy expands. Congestion results when traffic demand approaches or exceeds the available capacity of the system. In 2007, congestion, in terms of wasted time and fuel, cost about $87.2 billion in 439 urban areas, compared to $87.1 billion (in constant dollars) in 2006 (Schrank and Lomax, 2009). Since traffic volume will continue to grow rapidly, congestion, usually considered as a metropolitan problem, will become increasingly common in small cities and even rural areas(FHWA, 2005; Schrank and Lomax, 2009). To relieve congestion, various measures, such as tolling, congestion pricing, information provision, and high occupancy vehicle lanes/facilities, are proposed and being implemented throughout the United States. Among these measures, Intelligent Transportation Systems (ITS) serve as one of the most effective ways to tackle the congestion problem. For example, a combination of ITS technologies in Detroit, Michigan increased average vehicle speeds by 5.4 mph, decreased trip times by 4.6 minutes, and reduced commuter delay by 22 percent (FHWA, 2005). ITS plays an important role in improving transportation system efficiency and traveler convenience (Zhou, 2004).

The successful implementation and function of ITS architecture require constantly-monitored network traffic conditions. Based on the traffic conditions, sometimes a number of alternative schemes need to be evaluated for effective traffic management (Hellinga, 1994). For example, signal timing needs to be adjusted in correspondence to current traffic conditions. These evaluations usually need dynamic traffic demand in the form of origin-destination (OD) matrices (Bottom, 2000; Antoniou et al., 2006). Therefore, OD trip tables are important for these on-line traffic management applications. In addition, reliable forecasts of traffic demand and subsequently predicted network traffic conditions are critical to provide route guidance instructions as well as en-route diversion suggestions (Hellinga and Van Aerde, 1998). The inability of providing high quality OD demand estimates limits the potential for ITS deployments to alleviate traffic congestion and enhance mobility in urban networks (Zhou, 2004).

Various measures can be used to obtain OD demands for different application scenarios. For offline planning purposes, OD trip information can be obtained from direct interviews and/or surveys. Household, destination and roadside surveys are typically used in transportation planning analysis (Ortuzar and Willumsen, 1994). The survey method provides valuable samples about the detailed travel activities of each trip-maker, such as the origin and destination, the mode used and the travel time (Zhou, 2004). However, this method cannot satiate the needs of online applications, which requires dynamic demand information. Fortunately, due to the emerging traffic surveillance technology, real-time monitoring of traffic networks is possible and provides reliable data for deriving the OD demands. Traffic surveillance technologies, based on their functionality, can be categorized into three types, namely point, point-to-point and area wide (Antoniou et al., 2006). Point sensors include the widely-used loop detectors, which provide volume, occupancy, and speed measurements (Nihan et al., 2006). Point-to-point sensors, such as Automatic Vehicle Identification (AVI) systems, make the tracking of vehicles possible. More importantly, this technology provides direct information regarding OD demands (Cascetta and Marquis, 1993; Antoniou et al., 2006). Area-wide technologies, including airborne sensors, are promising technologies that are still being developed (Antoniou et al., 2006). The data retrieved from the first two categories of sensor technology have already been used as source data to estimate OD demands. In addition to OD estimation, prediction of OD demands is also beneficial within the context of ITS applications. Detailed discussion of these two problems will be provided in the following sections.

1.2. Problem Statement

As mentioned earlier, dynamic OD matrices of an extended period of time are almost impossible to acquire through direct measurements, interviews or field surveys. But by using traffic measurements such as traffic counts, turning movement counts, and other data from various sources, one reasonably estimate these unknown OD matrices. Two distinct but related kinds of problems need to be discussed, namely the dynamic OD estimation problem and OD prediction problem. Note since for Interstate 66, the object of this study, only loop detector data is available. From here on, all the data refers to loop detector data unless specified otherwise.

The ***Real-time OD estimation*** problem is defined as follows. Given link traffic counts for multiple time intervals $t = 1,2, \ldots, N$ and historical OD flows, estimate OD matrices for vehicles departing during each time interval t *until the current time interval* (Ashok, 1996). In comparison, the ***OD prediction*** problem is to forecast the number of vehicles departing at *future time slices* (Ashok, 1996; Camus et al., 1997).

Since OD matrices are the direct input for traffic assignment problem, another definition of the OD estimation problem is based on traffic assignment which specifically refers to the procedure that assigns a certain OD matrix to a specific road network based on given link performance functions (Sheffi, 1985). By exploitation of traffic assignment, one thus obtains the traffic volumes on each link of the transportation network to which traffic is assigned. This relationship between OD demand, flow pattern, and assignment procedure can be expressed by a functional form as follows (Sheffi, 1985; Yang et al., 1992):

$$x = \text{assign}(q) \tag{1.1}$$

where x is the flow pattern resultant from the assignment procedure applied to OD matrix q. The definition based on traffic assignment states that the OD estimation problem can be defined as the inverse problem of traffic assignment: find an OD matrix q that produces identical flow pattern x in the transportation network (Abrahamsson, 1998). In addition to the observed flow information, sometimes, prior knowledge of the unknown OD may be obtained from historical data or a sample survey. This prior information serves as an initial estimate of the true demand and is referred to as the target OD (Abrahamsson, 1998). Therefore, the information for OD estimation is twofold: observed flow pattern x and prior knowledge of the unknown demand. They can be integrated into a single estimation framework (Cascetta and Marquis, 1993) as follows.

$$\begin{aligned} &\min w_1 f_1(q, \bar{q}) + w_2 f_2(x, x^*) \\ &s.t. \\ &q \geq 0 \\ &x = assign(q) \end{aligned} \tag{1.2}$$

where $\bar{q}$ represents the OD demand previously known and x^*represents the estimated flow and function f_1,and $f_2$represents error measurement function. A typical form for $f_1$is mainly minimum-information type while f_2 typically takes Euclidean distance format (Abrahamsson, 1998).

Note that though the definition of OD estimation and prediction seems straightforward, the OD estimation and prediction problem are not easy in that there are still many complications.

- Firstly, the measurement errors and unavailability of traffic data for certain links or time slots always pose challenges to OD estimation and prediction. It is natural that observed traffic counts may not be available for all the links that constitute the network. Even if the traffic data is available, known and hidden errors adversely impact the OD estimation and prediction procedures.

- The assignment procedure used greatly impacts the flow pattern obtained. In the uncongested network case, usually, flow-independent proportional assignment such as all-or-nothing assignment is assumed (Yang et al., 1992). Though proportional assignment enjoys a computational advantage (Yang et al., 1992), the assumption of uncongested network does not always hold especially when disruptions such as incidents and extreme weather occur in the network.

- In addition, there are usually a large number of different OD matrices that may give the observed flow pattern. Hence, the OD matrix with maximum likelihood causing the observed traffic pattern is desired and subsequently, various techniques and methods have been proposed to deal with the selection of the "best" OD estimate.

- Finally, since loop detector data is the source for OD estimation and prediction, the real OD demands are unknown. Hence, the prediction method used should be capable of providing reliable estimates without knowing real OD matrices.

1.3. Objectives of the Project

The objectives of this project are stated as follows:

- Propose data correction procedures and evaluate their performance using the Interstate 66 (I-66) loop detector data.
- Model the traffic flow series using the Box-Jenkins time series technique and provide forecasts for future flows.
- Implement an OD prediction procedure for Interstate 66 using QueensOD and evaluate the possibility of real-time deployment of this procedure.

1.4. Methodology of the Project

The following measures are taken to tackle the problems aforementioned and achieve the objectives:

- Develop a data correction procedure to correct the erroneous data and fill missing data.

The choice of assignment principle does not pose a threat to OD estimation for a single freeway section due to the fact that only one route connects each OD pair.

- Use the QueensOD model for OD estimation. The model adopts the concept of least squares regression and attempts to minimize the sum of squared error between the observed and estimated link flows. This model conforms to the Euclidean distance format of function f_2. The QueensOD model will automatically choose the best OD estimate.
- Construct a procedure that is used to predict future link flows, which then will be used for OD prediction. In other words, instead of predicting OD demands directly, the future link flows are predicted and serve as the basis for "OD estimation" for future time intervals.

1.5. Organization of the Report

The project is divided into five chapters. After the introductory Chapter 1, Chapter 2 deals with the data correction issue by proposing two new data correction procedures and comparing them to a commonly-used procedure. Chapter 3 models the traffic flow series using the Box-Jenkins time series technique and provides forecasts for future flows. Chapter 4 applies QueensOD to use the predicted flows for OD prediction. The final chapter provides the concluding remarks.

CHAPTER 2 DATA PREPARATION

This chapter deals with the data preparation for OD prediction by addressing the problem of correcting single station loop detector data.

2.1. Introduction

As a result of congestion, U.S. urban travelers spent 4.2 billion more hours traveling and purchased an extra 2.8 billion gallons of fuel than the previous decade, which resulted in a monetary cost of $87.2 billion in 2007 (Hellinga, 1994). To help relieve congestion, Advanced Traveler Information Systems (ATIS), supported by the national Intelligent Transportation System (ITS) infrastructure, are implemented. ATIS offers a variety of benefits including decreasing travel time by 4% to 20% or more in severe congestion (Antoniou et al., 2006). Specifically, ATIS provides route and decision guidance with sufficient detail about possible alternate routes, travel times and locations in a timely and cost-effective fashion (Antoniou et al., 2006). To disseminate proper and useful advisory messages, one of the critical steps is collecting relevant data, which includes data synthesis and data translation (Antoniou et al., 2006). Moreover, the accuracy and reliability of the collected data greatly impacts the user-benefit of the ATIS (Abrahamsson, 1998; Bottom, 2000). Travelers may encounter worse conditions if they follow advisory messages that are based on erroneous data, which fail to reflect the real traffic conditions. The requested data, among other measurements, include volume and occupancy which can be readily retrieved from inductive loop detectors, an important data source for ATIS applications (Institute of Transportation Engineers, 2000). However, it is well known that loop detector data is prone to errors (Peeta and Anastassopoulos, 2002; Vanajakshi and Rilett, 2004). Therefore, it is essential to efficiently and effectively detect, diagnose and correct erroneous data for use in ATIS applications, travel time estimation and prediction, origin-destination demand estimation, and incident detection which rely on loop detector data streams (Fernandez-Moctezuma et al., 2007).

Loop detectors are usually installed on each freeway lane and parallel detectors constitute a detector station. Loop detectors are presence-type detectors, which sense the presence and passage of vehicles over a short section of roadway (May, 1990). The sensor is activated when a vehicle enters the detection zone and remains so until the vehicle leaves the detection zone. This activated duration (vehicle occupancy time) is the time for a vehicle to travel the length of the detection zone plus the vehicle's length (Daganzo, 1997). Typical measurements obtained from loop detectors are occupancy, volume, and speed at a pre-defined time resolution. Occupancy is a surrogate for density and volume can be easily transformed into hourly flow (May, 1990). Therefore, loop detector data essentially provide microscopic (time headways and vehicle speeds) and macroscopic characteristics of traffic flow (May, 1990), which can then be used to derive travel times for ATIS applications (Antoniou et al., 2006).

Issues associated with loop detector data are erroneous recordings and missing data, caused by improper installation, communication malfunction, and/or wire failures (Bikowitz and Ross, 1985). Current data collection is usually continuous, which makes equipment malfunctions more likely than if occasional collection techniques were used (Vanajakshi and Rilett, 2004). Continuous data collection leads to large amounts of data, which makes manual detection and correction techniques impractical. Therefore, efficient, automated procedures for detecting and correcting errors are needed.

Error detection and correction algorithms typically use historical data from the same detector, data from neighboring detectors, or a combination of both. Depending on the data used, a detection and/or correction procedure can be classified into two general categories: temporal or spatial. "Temporal" refers to procedures that rely on the historical data of the same detector (e.g. (Peeta and Anastassopoulos, 2002)), while "spatial" indicates approaches that use data from the detectors within the same station. Some approaches such as in Smith and Conklin (2009) combine spatial and temporal information. Spatial correction procedures (e.g. (Chen et al., 2003)) use historical data to derive the relationships between in-

station detectors for subsequent correction, but are categorized as spatial since correction is based on in-station detector data.

To aid with the selection of correction procedures for non-incident conditions, the performance of four different correction procedures which belong to the two general categories aforementioned are compared in a comprehensive manner using field data in the context of a single station. Among the four approaches are two proposed spatial correction procedures. One uses kernel regression to derive station flows based on lane use percentages and the other modifies an existing approach (i.e. (Chen et al., 2003)) by using lane flow to estimate station flow (whereas Chen et al. used one lane's flow to predict another lane's flow). These two new procedures are evaluated along with Smith and Conklin's (2002) approach and the time of day historical mean approach. The comparison is conducted from two aspects, namely systematic evaluation and random-error evaluation. The systematic evaluation is characterized by a completely deterministic error configuration while errors are introduced in a stochastic fashion for the random-error evaluation. The results of the comparison are analyzed to identify the most accurate method(s) to correct a large dataset and obtain accurate station flow estimates. In addition, recommendations for practical implementations are drawn from the comparison results.

The remainder of the paper contains six sections. Section 2 reviews earlier studies about loop detector data error detection and correction. Section 3 describes the characteristics of the dataset used for the subsequent comparison study. Section 4 presents the correction methods, followed by the experimental procedures in the fifth section. The results of the comparison are provided in Section 6. The last section presents conclusions and future directions.

2.2. Previous Studies of Loop Detector Data Error Correction

Error identification is necessary before any correction measure can be applied. Though this study does not deal with error detection, the detection methods are discussed for completeness and some studies jointly address the detection and correction of errors.

As mentioned earlier, techniques for detection and correction of errors in loop detector data fall into two general categories, namely temporal and spatial methods. But additional techniques based on checking for impossible combinations of traffic flow characteristics are also used to detect errors in loop detector data. Impossible combinations of traffic flow parameters can be identified by volume-to-occupancy or flow-to-occupancy ratios (Jacobson et al., 1990; Cleghorn et al., 1991). Turner (2004) later incorporated speed into combination checking. Chen et al. (2003) checked for impossible combinations over an entire day, rather than a smaller time period, and detected implausible time series of traffic measurements. Another widely-used error checking technique is to compare volumes or flows, occupancies, and speeds with specific thresholds (Weijermars and Van Berkum, 2006; Payne and Thompson, 2007).

Temporal detection and correction procedures were implemented by Chen et.al (1987) and many Intelligent Transportation Systems in Europe (Turner, 2004). Park (2003) constructed a multi-variant screening methodology to identify outliers of traffic data based on a variant of Mahalanobis distance. Ishak (1994) developed an algorithm using fuzzy theory that clusters the input space of the parameters speed, occupancy, and volume into regions of highly concentrated observations based on the normalized Euclidian distance. Each observation's level of uncertainty is then measured with one parameter, which is used for data screening to identify potential errors. Later this work was improved using probability theory and applied to real-time situations (Bhattacharjee et al., 2001; Alibabai and Mahmassani, 2008). Nihan (1997) modeled occupancy and flow time series as Autoregressive Moving Average processes (ARMA) and predicted values in the near future for the same station. Other efforts relying on time-series techniques include Maier et al. (1998), who applied robust estimation techniques to the time-series of flow, occupancy and speed measurements to estimate the corrected measurements. Using a spectral-domain time-series technique, Peeta and Anastassopoulos (2002) developed a Fourier-Transform based

algorithm that considered traffic characteristics as a time-dependent function and distinguished abnormal data caused by incidents from those due to detector malfunctions and subsequently corrected erroneous data. One of the major concerns of temporal detection and correction is that traffic flows could deviate from historical values due to special circumstances, such as incidents, instead of malfunctioning detectors (Weijermars and Van Berkum, 2006).

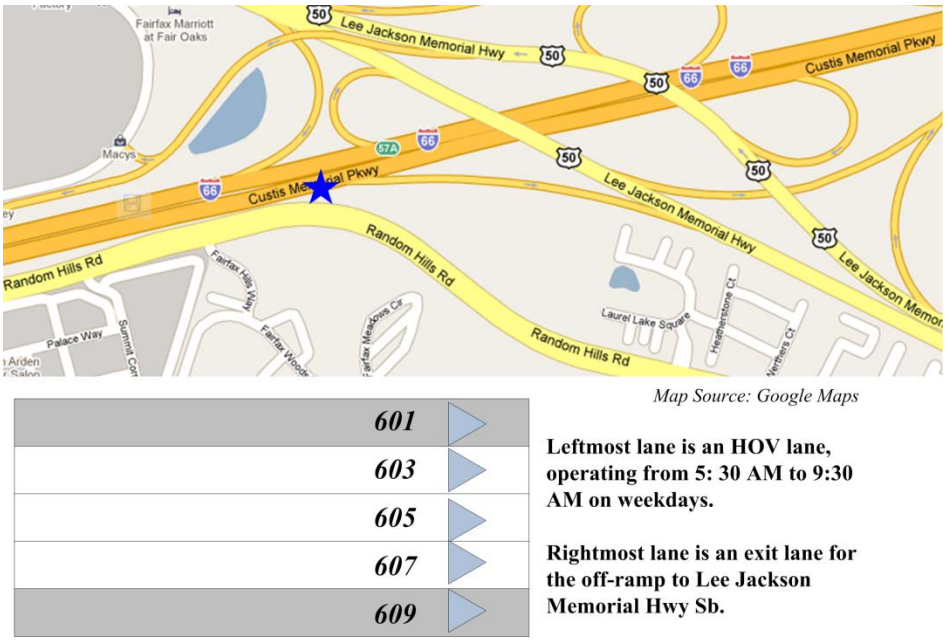
Spatial detection and correction procedures exploit the relationships between parallel detectors at a given station and/or upstream and downstream detectors. Kwon et al. (2004) developed an algorithm for detecting configuration errors based on the assumption that spatially close detectors should show similar temporal patterns if no major disturbances, such as freeway interchanges, exist in between. Chen et al. (2003) identified a strong correlation among neighboring loop detector measurements and modeled this correlation with a series of linear regression models. For a specific detector, a linear regression model was constructed using the data from this detector and one of its neighbors. The final detector flow estimate was the median of all the pair-wise estimates. Detection and correction based on the vehicle conservation principle compare traffic flow measurements between upstream and downstream locations. For example, Vanajakshi et al. (2004) compared cumulative volumes at each detector location and then corrected discrepancies using a constrained non-linear optimization procedure. Kikuchi et al. (1999) conducted another network-wise static data correction study where they applied fuzzy-optimization to adjust the flows of each link to achieve consistency. One potential deficiency of the spatial correction procedures is the inability to capture temporal variation of traffic throughout a typical day, which is considered to be critical in data correction (Abrahamsson, 1998).

A correction procedure that relied on both temporal and spatial information of the detectors is the one proposed by Smith and Conklin (2009), which used lane distribution patterns to derive the missing data values. They reported an average error less than 10% for almost all the test cases. This method will be discussed in detail in later sections and used in the comparison study.

From the array of previous approaches, Smith and Conklin's (2002), the time of day historical mean, and our modified version of Chen et al's (2003) approach were selected for comparison, along with our kernel regression approach, based on ease of implementation and reported correction accuracy. The variant of Chen et al.'s (2003) approach constructs a linear relationship between the detector and the station flow instead of between station-mates. In this study, the adoption of a kernel regression technique is inspired by its applications in traffic flow prediction (e.g. (Hellinga, 1994; Chen et al., 2008)). The main strength is its non-parametric nature, which requires no prior assumptions regarding the model structure (Takezawa, 2006). It is used to model the relationship between lane use percentage and lane flow and subsequently to correct erroneous flow readings. The details of the models are presented in subsequent sections.

2.3. Dataset Description

The Virginia Department of Transportation (VDOT) has installed and maintains 91 detectors grouped into 34 stations on eastbound Interstate 66 (I-66E) between Manassas and Falls Church, VA. Provided by VDOT, detector data and incident records from 2008 were used for this study. From this dataset, the data for the station near the Lee Jackson Memorial Highway interchange was extracted. This station was selected because it relates to other ongoing research efforts. The station has five detectors, which occupy each lane. The detector configuration and the location of the station are illustrated in Figure 1.



Map Source: Google Maps

Figure 1 Station Location and Detector Configuration

Each detector is expected to report speed, occupancy and volume every minute and 1440 records of station flow should exist for a day if no detector malfunctions occur. The real-time hourly flow is the volume reading multiplied by 60. The working status of the detectors is recorded by an indicator, which has three possible values, 0, 1, or 2. A "0" indicates the detector is not functional; a "1" is considered to be the flag for possible erroneous readings. Only readings from a detector with a status of "2" are considered correct.

When the status is 0 or 1, the readings are considered to have known errors (type 1). In addition to these errors, the detector status is not correctly reported at all times. Type 2 errors may include unreasonably high flows; considering that the detectors are individual-lane based, the hourly flow obtained from 1-minute data for a freeway lane cannot exceed 5000 vehicle/hr (Payne and Thompson, 2007). Hence, the detector status is set to 0 for those detectors reporting hourly flows greater than 5000 vehicle/hr for the corresponding lane. Type 3 errors, missing data, are common. Sometimes, the readings of individual detectors are missing, and at others no detectors report values for an extended duration, i.e. more than 10 minutes. Station flow is the sum of the individual detector flows, and thus, is correct only when all five detectors report correct values. In this study, all three error types are converted to missing-data errors since data retrieved from detectors with a 0 or 1 status are not reliable. Hence, in this study, the data correction is equivalent to estimation of missing data points and thus the two terms are used interchangeably. In addition, this conversion simplifies the implementation of the correction procedures.

To ensure a fair comparison, three datasets were carefully chosen and constructed for the study. The first dataset is the correction target. The data for Tuesday, April 22, 2008 during 4:00 AM to 10:00 PM was selected for correction technique comparison, similar to previous studies (Peeta and Anastassopoulos, 2002; Payne and Thompson, 2007; Schrank and Lomax, 2009). The second dataset is the correction base dataset, which contains the data for the three months prior to the target day (01/23/2008 to 04/20/2008) with all holidays excluded. The third dataset is a subset of the correction base dataset. It contains the data for Tuesdays in the correction base dataset. This subset serves as the correction base dataset for temporal correction that will be described in the next section. Additionally, the erroneous data points and those for durations in which an incident occurred are also eliminated from the base dataset, as in Peeta and Anastassopoulos (2002). The elimination of incident-related data is based on incident location and duration. The relevant incident location is documented as "I-66E at Lee Jackson Memorial" in the

incident records. The data for the reported duration of the incident was eliminated from the base dataset. The reasons why April 22, 2008 was chosen are as follows.

- If all detectors had reported correct readings, there would have been 1080 correct station flows for this time duration. The chosen day is among those with the highest number of correct station flow readings - 937 out of 1080. These 143 (1080 – 937 = 143) incorrect station flows were attributed to 259 malfunction detector readings. The detector malfunctions occurred primarily between 7 and 8 AM and these observations were not part of the correction procedure evaluation.
- In addition to data quality consideration, the correction base dataset does not involve seasonal shift, which may lead to a major change in the traffic pattern due to school holidays.
- No incidents occurred on this day for this particular location or within 5 miles upstream or downstream. The non-incident condition is explicitly required by the correction procedure proposed by Smith and Conklin (2009).

2.4. Data Correction Procedures

The four data correction procedures are illustrated within the context of the target dataset. First the basic techniques are presented and then their adaptations for this study are discussed.

2.4.1.Temporal Correction (TC)

The temporal correction (TC) applied for this study is the Time-of-Day (TOD) historical mean (Cascetta and Marquis, 1993; Daganzo, 1997; Abrahamsson, 1998). Let d denote the number of different days within the temporal base dataset (i.e. subset of the base dataset - all Tuesdays in the base dataset). The corrected flow for detector i at time t, $\hat{f}_t^i$ is calculated by eq.(2.1):

$$\hat{f}_t^i = \frac{\sum_{j=1}^{d} y_t^i(j)}{d} \tag{2.1}$$

where $y_t^i(j)$ denotes the flow reading for detector i at time t on day j. The temporal correction procedure takes the arithmetic average of the flows at time t of all days available in the base dataset. If a detector is functional, no correction is needed and the corrected flow $\hat{f}_t^i$ is equal to the reported flow f_t^i. The station flow is the summation of all the estimated detector flows and mathematically expressed as Eq.(2.2).

$$\hat{q}_t^{TC} = \sum_i \hat{f}_t^i \tag{2.2}$$

where $\hat{q}_t^{TC}$ denotes the corrected station flow (vehicles per hour) for time t using the temporal correction.

2.4.2.Spatial Correction using Linear Regression (LR)

Unlike Chen et al. (2003) who used relationships between flows of individual detectors, in this study, a series of linear regression models are constructed to estimate the *station* flow from functional detectors. Note that the time span of the correction target dataset is 4:00 AM to 10:00 PM, which includes the HOV-effective duration (5:30 AM to 9:30 AM). To deal with traffic pattern shifts due to the HOV restriction, two sets of regression models are built. One is for the HOV period and the other is for the regular period. For notational simplicity, the two sets of regression models are written in a unified format. Specifically, an estimate of the station flow $\hat{q}_t^i$ at time t based on the flow f_t^i of functional detector i at time t is obtained through the linear regression model expressed in eq.(2.3)

$$\hat{q}_t^i = b_1^i f_t^i + b_2^i + e,\ i = 1, 2, 3 \ldots 5 \tag{2.3}$$

Note that only readings from functional detectors are used. The regression coefficients b_1^i and b_2^i are estimated using the ordinary least squares method. For the models when the HOV lane is activated, only the data points taken during the HOV activation in the base dataset are used to estimate the coefficients for HOV models while the data points for the regular period in the base dataset are used to estimate the

coefficients of regression models for the regular period. The final estimated station flow at time t was computed as the arithmetic average of all available estimates of station flow using Eq. (2.4).

$$\hat{q}_t^{LR} = \frac{\sum_i \hat{q}_t^i}{n} \tag{2.4}$$

In Eq. (2.4), n denotes the number of functional detectors at a given time t. For a given time t, it is essential to determine whether this time stamp belongs to the HOV duration or not and choose the set of models accordingly. If three detectors, for example, are functional at this time, then an estimate of station flow can be derived from each detector flow reading. The final estimated station flow is the arithmetic mean of these three station flow estimates. It can be seen that the linear regression procedure requires that at least one detector be fully functional in order to derive valid estimates of the station flow at a specific time. If no detectors report correct flow readings, the station flow is estimated using the temporal procedure $\hat{q}_t^{TC}$, i.e. the temporal correction approach serves as a default technique when all detectors fail.

2.4.3.Spatial Correction using Kernel Regression (KR)

Under different traffic conditions (congestion levels), drivers use different lanes, creating lane-to-lane flow variability in multi-lane freeway sections (Carter et al., 1999; Amin and Banks, 2005). The lane use pattern depends on the location and flow conditions (Nihan et al., 2006). In a study of three-lane freeways under different flow conditions, Pignataro (1998) found that (1) under conditions of low flow, the center lane carried the majority of the flow while the median lane carried the lowest percentage, (2) as the total volume increased, flow percentages in the shoulder and center lanes dropped, while the percentage of total flow in the median lane increased, and (3) when the flow approached section capacity, the percentages of flow in the median and center lanes were equal, at approximately 37 percent. Other studies (May, 1990; Carter et al., 1999; Fernandez-Moctezuma et al., 2007) also reported relationships between the lane use pattern and flow conditions.

This study proposes a new spatial correction procedure that exploits the relationship between lane use pattern and flow conditions. A functional detector i at time t reports flow f_t^i. The lane use percentage p_t^i is defined as in Eq.(2.5).

$$p_t^i = \frac{f_t^i}{q_t^i} \tag{2.5}$$

If the lane use percentage p_t^i corresponding to the flow reading is known, the station flow can be estimated by eq.(2.6).

$$\hat{q}_t^i = \frac{f_t^i}{p_t^i} \tag{2.6}$$

However, no parametric function exists that can be used to describe the relationship between lane use percentage and lane flow. Therefore, kernel regression was used to obtain the empirical relationship between the lane flow and lane use percentage for each lane.

In parametric regression with a single regressor, the equation takes the form $y = F(x) + e$ like Eq.(2.3), where $F(x)$ is a smooth function with known shape described by a number of parameters. The m observations (x_j, y_j) $(j = 1,2,3,\ldots,m)$ are used to estimate the unknown parameters and subsequently the fitted value $\hat{y}_j$ is given by $F(x_j)$. For any given x_i, one can always get the estimated $\hat{y}_i$ using the constructed parametric relationship F. Similarly, the output of kernel regression is also a relationship that can be used to estimate $\hat{y}_i$ based on a provided x_i. The only difference lies in the fact that kernel regression does not make assumptions about the shape of this function $F(x)$ (Bowman and Azzalini, 1997).

The underlying logic of kernel regression is that every observation other than the point of interest (x_i, y_i) is used to derive the fitted value $\hat{y}_i$ and the observations with the most information about $F(x_i)$, or $\hat{y}_i$, should be those close to this point of interest (x_i, y_i) (Bowman and Azzalini, 1997; Takezawa, 2006). Note that the closeness of an observation (x_j, y_j) (j = 1,2,3,…,m and $j \neq i$) to the observation in question (x_i, y_i) is measured by the distance on the horizontal axis, which is calculated by $|x_j - x_i|$. The immediate neighboring observations, those with smaller distance $|x_j - x_i|$, contribute more to deciding $F(x_i)$ than distant ones. Therefore, a weight w_{ij} is introduced to measure the magnitude of contribution of an observation j to the fitted value $F(x_i)$. Subsequently, $F(x_i)$ is a weighted-average of all y's of the observations (x_j, y_j) (j = 1,2,3,…,m and $j \neq i$). Since the observations far from the observation of interest (x_i, y_i) should receive little or no weight, a decreasing function is desired to assign weights to each observation (x_j, y_j) based on the distance $|x_j - x_i|$ ($i \neq j$). After the weights for the value y_j are determined, $F(x_i)$ is computed as the weighted average of all the y_j's. After this process is repeated for each observation, the shape of the function $F(x)$ can be described by a series of derived points (x_i, $F(x_i)$) for i = 1,2,3,…,m where m is the total number of observations. The choice of the weighting function should not influence the relationship being captured by the kernel regression. Therefore, an additional parameter bandwidth, denoted by h, is introduced to rescale the horizontal distance so as to minimize the effect of different choice of weighting function. Hence, the ultimate measure of distance, used to decide the contribution of an observation to the calculation of fitted value, is the horizontal distance scaled by the bandwidth, i.e. $|x_j - x_i|/h$.

In this study, the five detectors are treated separately by constructing a kernel regression based relationship between individual lane flow and lane use percentage for each of them. Suppose m pairs of lane flow f_j and lane use percentage p_j (j = 1,2,3,…,m) for one particular detector are available in the base dataset which contains flows up to every minute of the past three months of 4 AM to 10 PM for every day. For a specific observation (f_i, p_i), a series of weights w_{ij} (j = 1,2,3,…,m and $j \neq i$) are assigned to each lane use percentage p_j selected out for this detector to determine fitted lane use percentage $\hat{p}_i$ based on the distance measure $|f_j - f_i|$. This step is repeated for every pair of lane flow and lane use percentage (f_i,p_i) for i = 1,2,3,…,m. The weighting function in this study is the one proposed by Nadaraya and Watson (2004). The weight for lane use percentage p_j was defined based on the difference between lane flow f_i and f_j as shown in eq.(2.7).

$$w_{ij} = \frac{K_h\left[\left(f_i - f_j\right)/h\right]}{\sum_{j=1}^{j=m} K_h\left[\left(f_i - f_j\right)/h\right]} \tag{2.7}$$

To simplify the notation, let $u = \left(f_i - f_j\right)/h$. Then $K_h\left(u\right)$ is a function of u and h is a positive-value bandwidth. The kernel function $K_h\left(u\right)$ used here is the standard Gaussian kernel, expressed by eq.(2.8).

$$K_h\left(u\right) = \frac{1}{\sqrt{2\pi}} e^{-u^2} \tag{2.8}$$

The equation proposed by Bowman and Azzalini (1997) was used to determine the optimal bandwidth for the standard Gaussian kernel. The optimal h was computed using eq.(2.9).

$$h_{\text{opt}} = \left(\frac{4}{3m}\right)^{\frac{1}{5}}, \tag{2.9}$$

where m is the sample size of the base dataset selected for this particular detector. By using eqs.(2.7) to (2.9), the fitted data point $\left(f_i, \hat{p}_i\right)$ can be found and thus the relationship between lane use percentage and lane flow F can be constructed for this detector. By applying this procedure to all five detectors, then for any lane flow, the corresponding lane use percentage can be estimated through interpolation based on the relationship F.

Once the lane use percentage is determined, the station flow is estimated using eq.(2.6). Similar to the LR-based spatial correction, the final estimate of the station flow at time *t*, denoted by $\hat{q}_t^{KR}$, is the arithmetic mean of all the estimates from individual functional detectors.

The HOV restriction has a potentially large impact on the lane flow distribution. Therefore, the relationship between the lane flow and lane use percentage is treated separately for the HOV duration and regular period.

2.4.4.Lane Distribution Correction (LD)

The correction method proposed by Smith and Conklin (2009) exploits both the temporal and spatial information. Their approach is as follows. A historical lane use percentage $\bar{p}_t^i$ is derived from the correction base dataset. It is computed as the arithmetic mean of all the available lane use percentages for lane *i* at time *t* as shown in eq.(2.10):

$$\bar{p}_t^i = \frac{\sum_{j=1}^{d} p_t^i(j)}{d}, \tag{2.10}$$

where *d* denotes the number of different days within the base dataset and $p_t^i(j)$ is the lane use percentage of lane *i* at time *t* of day *j*, which can be calculated using eq. (2.5). This historical lane use percentage is used to estimate the station flow as indicated by eq.(2.11).

$$\hat{q}_t^i = \frac{f_t^i}{\bar{p}_t^i} \tag{2.11}$$

The final estimate of the station flow for time *t* is the arithmetic mean of all available estimates based on individual detectors. Specifically, it is written as eq.(2.12):

$$\hat{q}_t^{LD} = \frac{\sum_i \hat{q}_t^i}{n}, \tag{2.12}$$

where *n* is the number of functional detectors at time *t*.

2.5. Experimental Procedures

The experiment consists of two components, namely systematic evaluation and random-error evaluation. The systematic evaluation assesses the performance of each correction method under all possible error combinations. Since five different detectors are involved, the number of different error combinations is $2^5 - 2 = 30$. The two cases excluded are (1) all detectors are malfunctioned and (2) all detectors are functional. Figure 2 visualizes the specific configurations for each error combination. In the random-error evaluation, errors are introduced in a stochastic manner at different levels: 10, 20, 30, 40, and 50%.

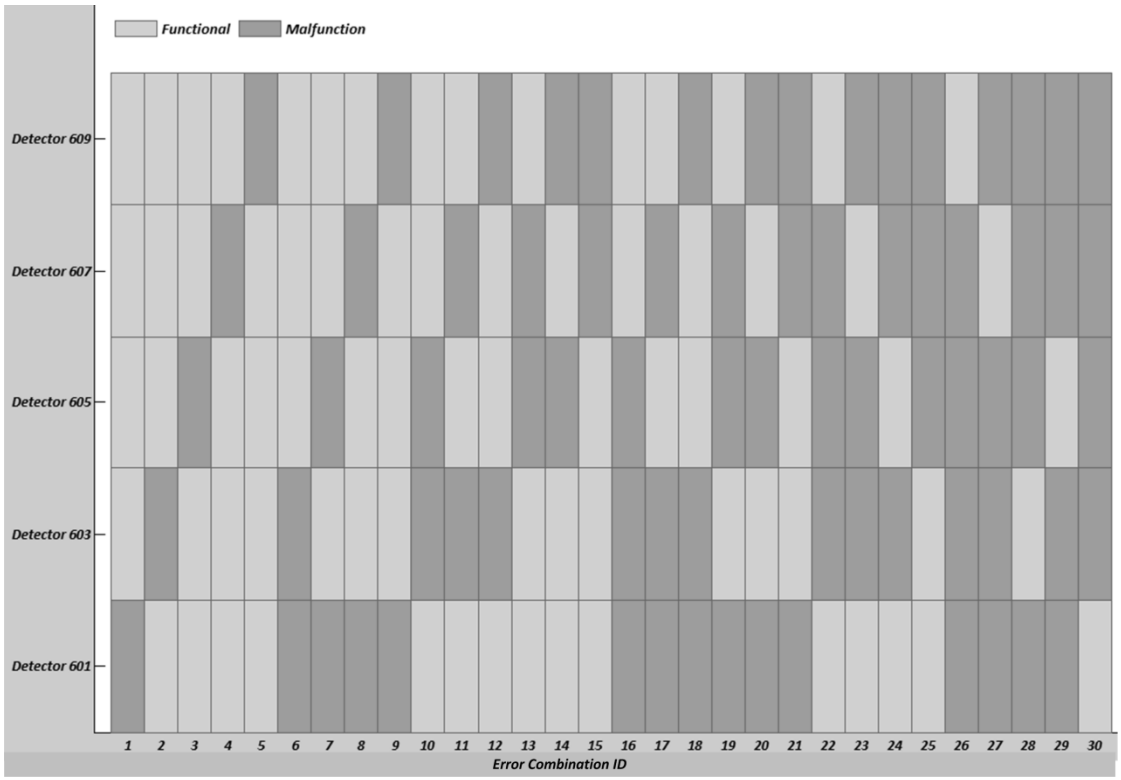


Figure 2 Error Configurations for Systematic Evaluation

The performance of the correction methods are compared using two widely-accepted measures of effectiveness (MOEs): Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). The MAE metric is defined as in eq.(2.13).

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T} | q_t - \hat{q}_t |, \tag{2.13}$$

where T represents the total number of records in the target dataset and q_t is real station flow and $\hat{q}_t$ is the estimated station flow. The MAPE metric converts the absolute quantity to a relative one, which is written as in eq.(2.14).

$$\text{MAPE} = \frac{1}{T}\sum_{t=1}^{T} \frac{| q_t - \hat{q}_t |}{q_t} \tag{2.14}$$

The systematic evaluation targets two periods, namely a peak hour and an off-peak hour, within the duration of 4:00 AM to 10:00 PM. The selected peak hour is from 8:00 AM to 9:00 AM while the off-peak hour starts at 2:00 PM and ends at 3:00 PM. For each error combination, the flow readings of the desired detectors are set to be missing for the entire hour.

Two systematic evaluations are conducted with different purposes. In the first systematic evaluation (***SysEval-I***), the KR and LD approaches adopt all available estimates of station flows to derive the final estimate of station flow for a given time. The results of the first systematic evaluation (***SysEval-I***) are analyzed by comparing the performance of correction procedures under different error scenarios so as to provide insights into the relative importance of each detector as to the accuracy of data correction methods. These insights were used to modify the LR and KR spatial correction methods. The second systematic evaluation (***SysEval-II***) is then conducted by incorporating the modifications. Based on the results of ***SysEval-II***, observations about their performances under different numbers of malfunctioning detectors are made. The procedures are also compared under each error scenario to identify the one that provides the best performance.

The systematic evaluation is followed by a random-error evaluation. As mentioned earlier, the target day contains a total of 5400 individual detector flow records or 1080 station flows. Errors are introduced into the target dataset in a random fashion at five different error levels ranging from 10% to 50%. The 10% error level means that 10% of the individual flow readings (i.e. 10%×5400 = 540 records) will be artificially set to missing values. Specifically, a random number between 0 and 1 was generated and assigned to each detector flow record. If the random number is less than the error level, then this record is chosen to be set to missing value. Recall that the target dataset contains 259 pre-existing incorrect detector readings. The actual error percentage could be less than the specified error level when the existing incorrect detector records are flagged as introduced error. For each error level, 10 replications are carried out by revising the random number seed so as to eliminate statistical bias. Each replication hosts a particular combination of the error configurations defined in the systematic evaluation.

The results of random-error evaluation are analyzed and interpreted from several angles as follows. The overall performance of each procedure will be evaluated and contrasted at different error levels. The corrected value is supplied by temporal correction if the other three procedures (LR, KR and LD) fail to produce an estimate of the station flow for a given time. Such correction failures of the three procedures aforementioned are assessed to provide insights as to the robustness of the correction procedures. Finally, the overall performance measure for the random-error evaluation is disaggregated by time of day to uncover the temporal performance variations.

2.6. Experimental Results

The two sets of linear regression models need to be constructed using the correction base dataset before the LR correction can be conducted. The regression models for the HOV period and the regular period are shown in Figure 3 and Figure 4, respectively. The linear relationship between the individual lane flow of detector 609 and the station flow for the HOV period is weak, which is evidenced by the relatively low adjusted R^2 metric 0.01. The same findings apply for individual lane flow of detectors 607 and 609 for the regular period. Their adjusted R^2 are 0.23 and 0.38 respectively. Using weak linear relationship leads to unreliable results. Therefore, for the HOV period, the final estimate of station flow is derived as the arithmetic mean of available estimates from the detector 601, 603, 605 and 607. For the regular period, only the available estimates based on individual lane flows provided by detectors 601, 603 and 605 are used for computing the final estimated station flow.

2.6.1.Results of Systematic Evaluation

The results of the first systematic evaluation (***SysEval-I***) are shown in Figure 5 and Figure 6. Note that the purpose of ***SysEval-I*** is to identify the relative importance of lane detectors with respect to the accuracy of correction procedure.

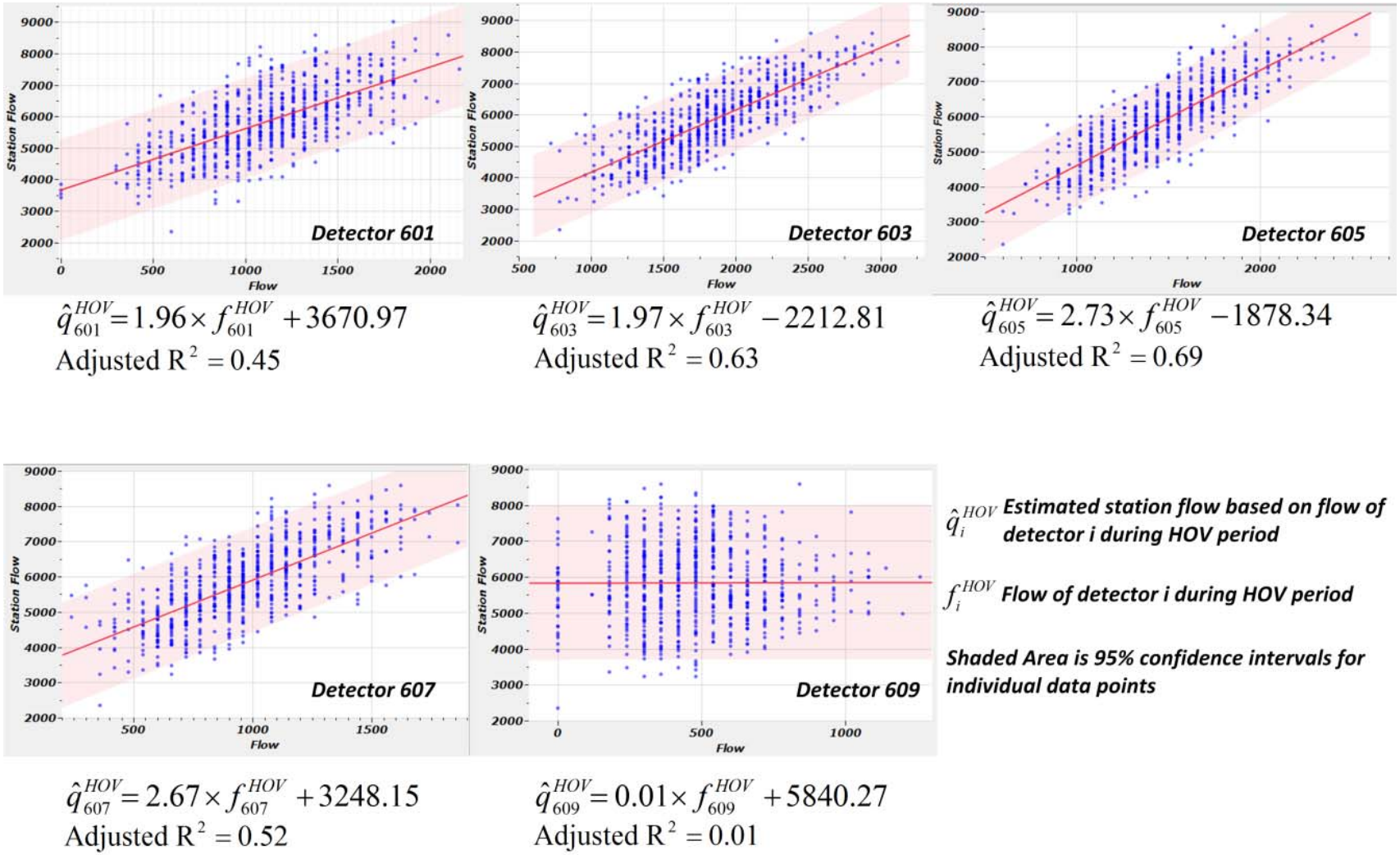
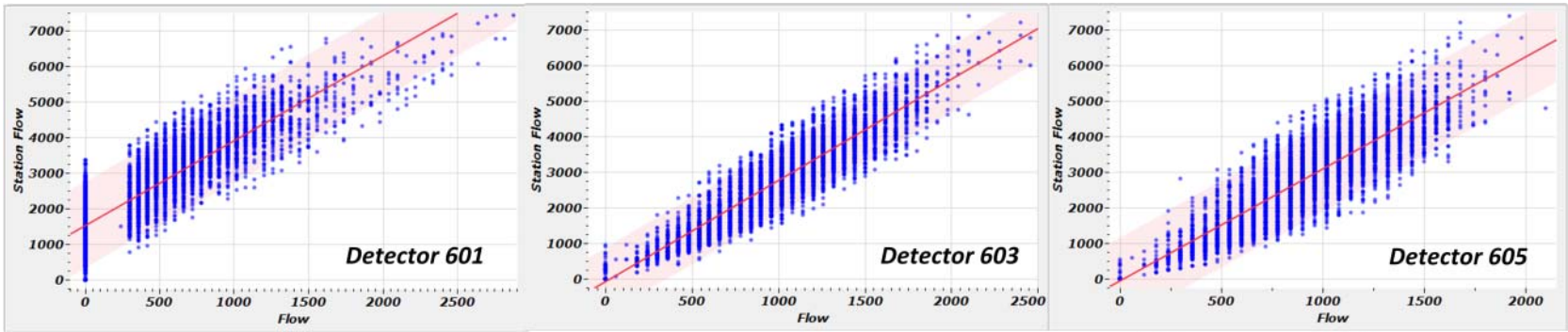


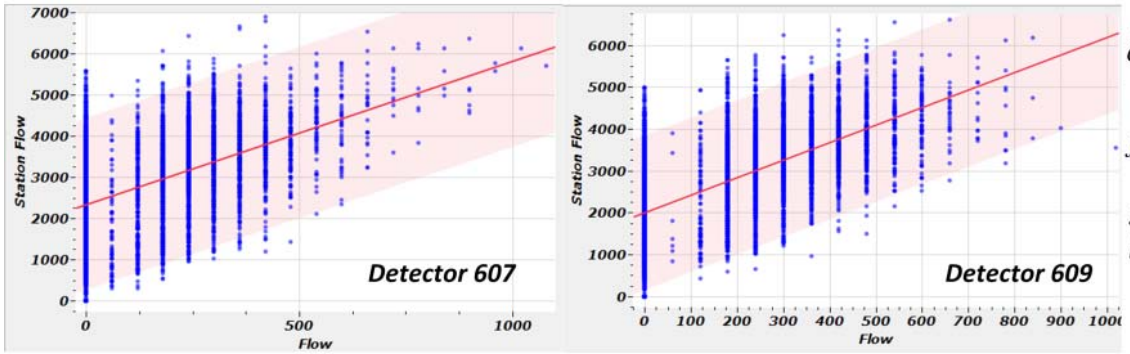
Figure 3 Linear Regression Models for the HOV Period



$\hat{q}_{601}^{REG} = 2.38 \times f_{601}^{REG} + 1543.21$
Adjusted $R^2 = 0.75$

$\hat{q}_{603}^{REG} = 2.84 \times f_{603}^{REG} - 65.14$
Adjusted $R^2 = 0.85$

$\hat{q}_{605}^{REG} = 3.15 \times f_{605}^{REG} - 47.55$
Adjusted $R^2 = 0.75$



$\hat{q}_{607}^{REG} = 3.46 \times f_{607}^{REG} + 2353.18$
Adjusted $R^2 = 0.23$

$\hat{q}_{609}^{REG} = 4.19 \times f_{609}^{REG} + 2006.42$
Adjusted $R^2 = 0.38$

$\hat{q}_i^{REG}$ ***Estimated station flow based on flow of detector i during regular period***

f_i^{REG} ***Flow of detector i during regular period***

Shaded Area is 95% confidence intervals for individual data points

Figure 4 Linear Regression Models for the Regular Period

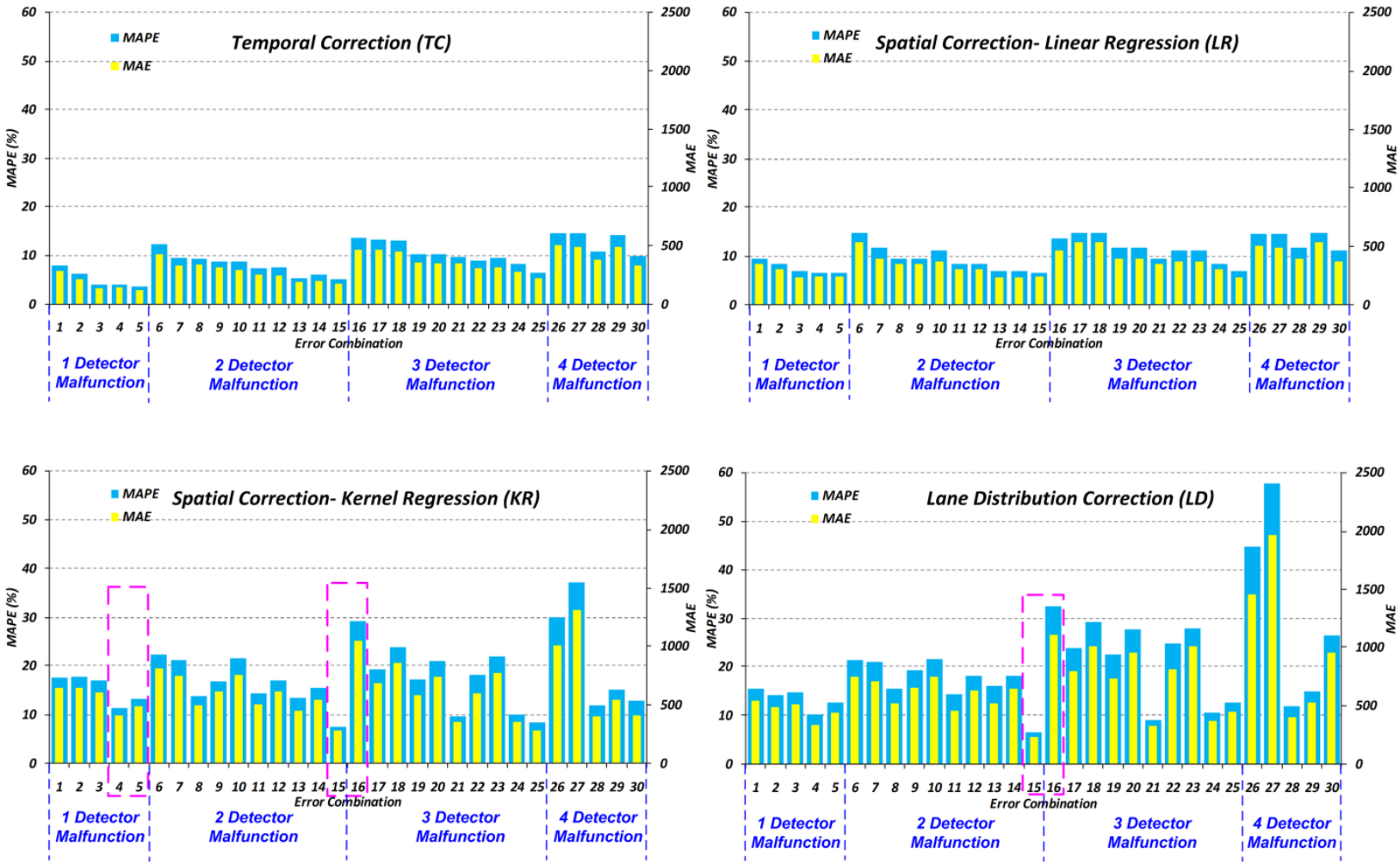


Figure 5 Results of SysEval-I for the Off-peak Hour

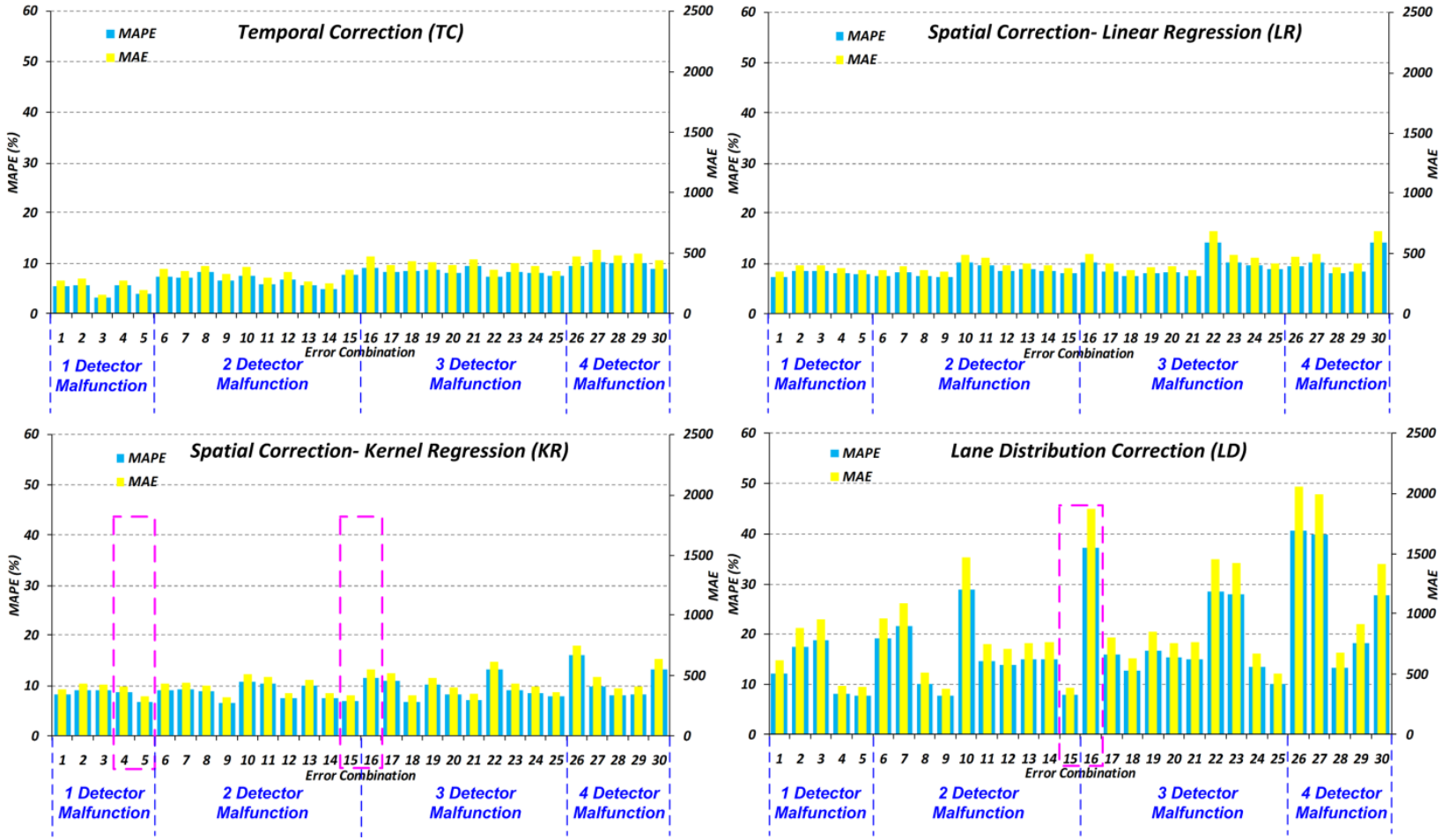


Figure 6 Results of SysEval-I for the Peak Hour

An important observation can be made that the MOEs were not uniform within a scenario group of the same number of malfunctioned detectors. Equivalently, the location of malfunctioned detectors impacts the accuracy of the correction procedure. (To simplify the narration, detector 601 malfunctioned is represented by the symbol "d601-X".) Specifically, for the off-peak period, the temporal correction method (upper left panel of Figure 5) achieved lowest MAPE and MAE for combination 5 (d609-X, see Figure 2). For the temporal correction method, the MAPE for combination 6 (d601-X&d603-X) is significantly higher than that for combination 5. It can be seen that the information from detectors 601 and 603 plays an important role in temporal correction. Similarly, the information from detectors 601 and 603 is important to all correction methods considered in this study. This observation is further substantiated by combinations 26 and 27 exhibiting high MAPE for all correction methods; for these two combinations, detectors 601, 603 and 605 malfunctioned.

Another interesting comparison is the one among error combinations in the off peak period 4 (d607-X), 5(d609-X), 15(d607-X & d609-X) and 16(d601-X, d603-X & d605-X). For the spatial correction method using kernel regression, the MAPE of combination 15 is even lower than combination 5 as well as combination 4. This indicates that more information does not necessarily lead to more accurate estimation. In fact, the addition of information from either detector 607 or detector 609, in the absence of another detector, decreases the effectiveness of the KR correction method. In addition, the MAPE of combination 16 is higher than combinations 4, 5 and 15, either of which at least has one functional detector of the detector group (d601, d603 and d605). A similar pattern is observed in the results of the peak period. The slight difference lies in the fact that the KR approach achieved the lowest MAPE (8.40%) for combination 25 (d605-X, d607-X & d609-X) for the off-peak period while the lowest MAPE 6.92% was achieved for combination 15 (d607-X & d609-X) during the peak period. The practical significance of this finding is that an adaption to the KR approach is needed. Specifically, during the off-peak hour, only available estimates of station flow based on individual lane flows provided by detector 601 and 603 are used. During the peak hour, the lane flows retrieved from detectors 601, 603 and 605 are used for final estimate of station flow.

A similar pattern can be found for the lane distribution correction method (lower right panel in Figure 5 and Figure 6). The MAPE for combination 15 is much lower than that for combination 16. In other words, the estimated station flows based on individual flows of the two right-most lanes are relatively inaccurate. This is because the right-most lane is the exit lane, whose flow pattern exhibits time-dependent as well as day-to-day variation. Based on this finding, the LD correction method is revised to use only available estimates from detector 601, 603 and 605 for deriving the final estimate of station flow.

The findings of ***SysEval-I*** are used to modify the KR and LD approaches to improve their performance and the systematic evaluation is then conducted again and labeled as ***SysEval-II***. Note that the intent of ***SysEval-II*** is understanding the number of malfunctioned detectors and identifying the best method for every error configuration. The results of ***SysEval-II*** are shown in Figure 7 and Figure 8.

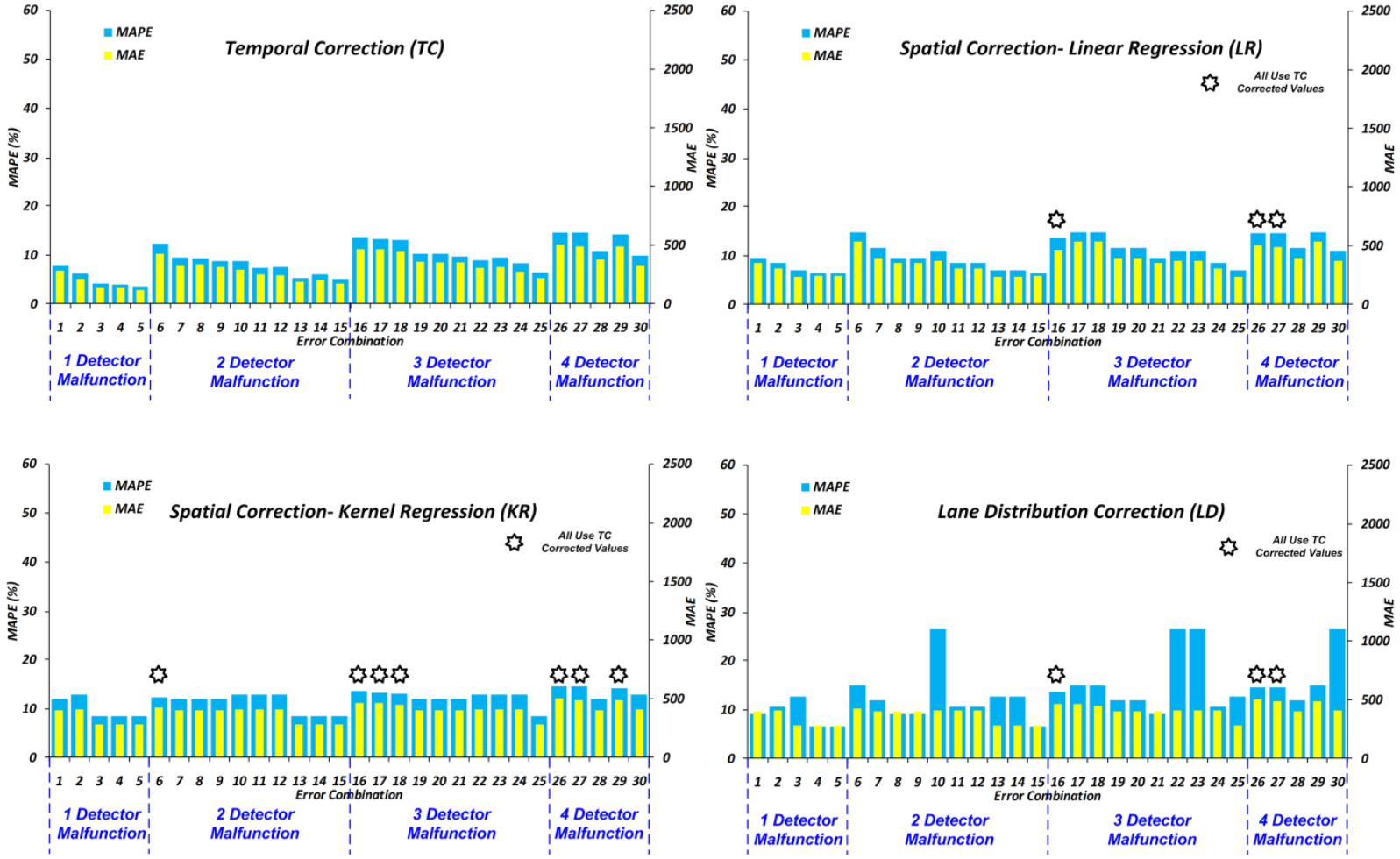


Figure 7 Results of SysEval-II for the Off-peak Hour

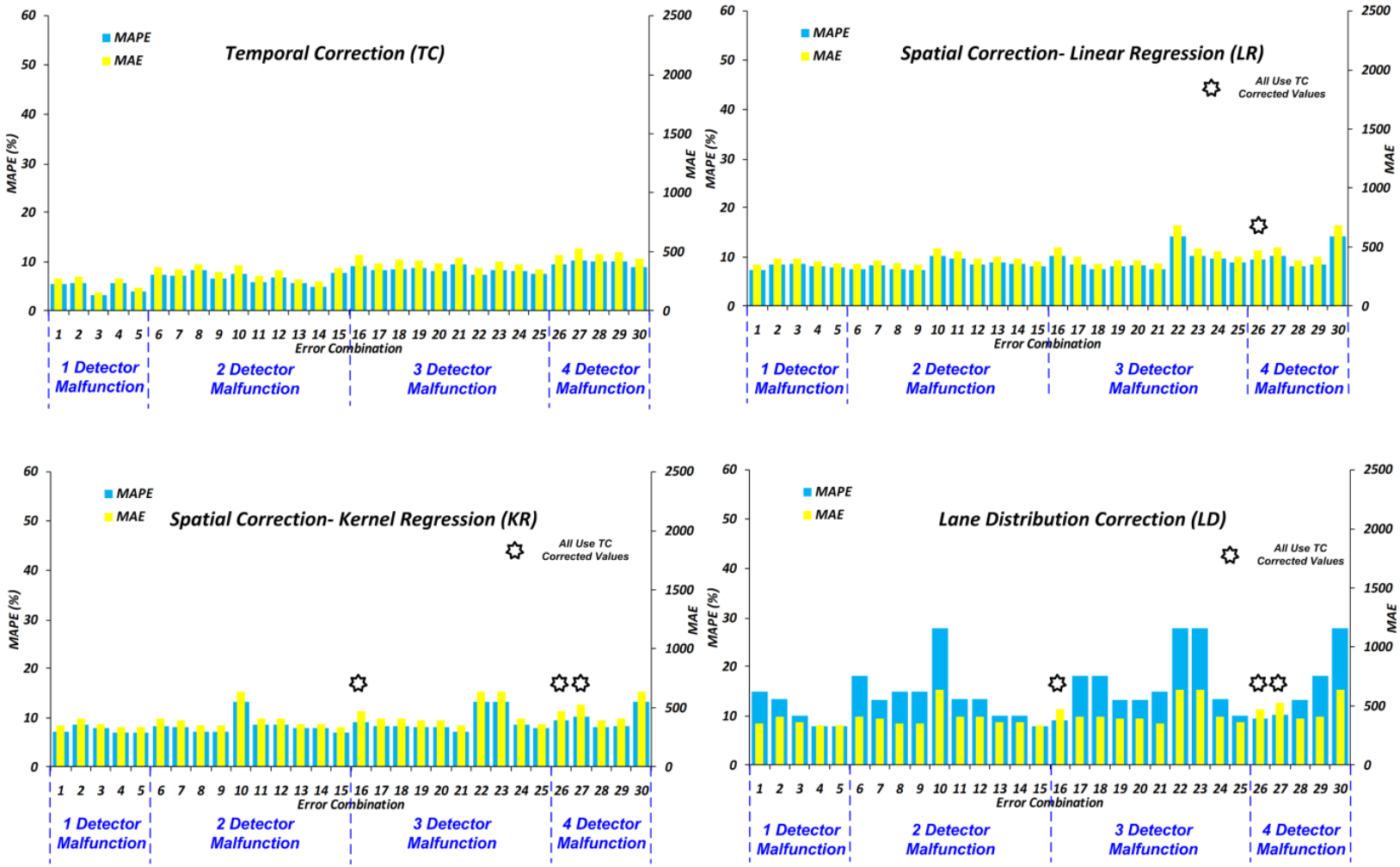


Figure 8 Results of SysEval-II for the Peak Hour

First, the improvement achieved by the modifications based on results of the previous systematic evaluation is confirmed by the obvious drop of the MAPEs of the LD approach for both the peak and off-peak hours as well as the decrease of the MAPEs of the KR approach for the off-peak hour. Therefore, it is once again reinforced that the relative importance of each detector is different for the four correction procedures.

Additionally, it is seen that the number of malfunctioned detectors influenced the performance of temporal correction and the LR spatial correction in a relatively more deterministic fashion than it did the other two approaches during the off-peak period. This is evidenced by the general upward trend of the MOEs when the number of malfunctioned detectors increased from one to four. However, this trend did not present itself in the performance of KR approach and LD approach. It thus follows that these two approaches are less sensitive to the number of malfunctioned detectors during the off-peak period. By a comparison between the off-peak period and peak-period, this upward trend was less apparent for the peak period than that of the off-peak period.

Table 1 highlights the lowest MAPE for each error configuration, which can be used to identify the correction procedure that provides superior performance. The procedure that provided the best performance is highlighted while the MAPE differences are calculated for the other three procedures. For example, temporal correction had the lowest MAPE (5.42%) for error combination 1 during the peak hour. The MAPE difference is "+1.79%" for the linear regression (LR) approach, which results in a MAPE of 7.21% (i.e. 5.42%+1.79%) for the same error configuration. For the off-peak hour, temporal correction (TC) presented dominant performance. The only exception is error combination 21 in which LD provided the lowest MAPE 9.04%. Note that the LR approach did perform very closely when TC is the best, which is proved by an average MAPE difference of 1.48%. For the peak hour, a similar pattern existed for the scenarios in which the number of malfunctioned detectors is less than 2 (before error combination 15). Specifically, temporal correction provided the most accurate correction results. When the number of malfunctioned detectors becomes three or more (beyond error combination 15), the KR approach provided competitive performance for several error configurations for the peak hour. Again, the LR approach is the second best with an average MAPE difference 1.95%. Hence, it can be seen that the linear regression approach exhibits robustness across different error configurations.

In addition, there are error combinations for which all correction methods exhibited higher error rates. For the off-peak hour, error combinations 17 and 18 resulted in MAPEs more than 13% for all procedures. The most critical reason is because detector 601 and 603, considered to be most important for all procedures, malfunctioned. For the peak hour, the similar combination is combination 22 (only detector 601 and 609 are working).

Table 1 MAPEs for the Different Error Configurations

Error Config.	*MAPE (%) 8:00 AM - 9:00 AM*				*MAPE (%) 2:00 PM - 3:00 PM*			
	TC	*LR*	*KR*	*LD*	*TC*	*LR*	*KR*	*LD*
1	*5.42*	*+1.79*	*+1.61*	*+9.52*	*7.77*	*+1.71*	*+4.18*	*+1.27*
2	*5.63*	*+2.82*	*+2.87*	*+7.75*	*6.20*	*+2.18*	*+6.56*	*+4.32*
3	*3.12*	*+5.45*	*+4.65*	*+6.87*	*4.00*	*+2.8*	*+4.4*	*+8.63*
4	*5.62*	*+2.38*	*+1.31*	*+2.19*	*3.95*	*+2.44*	*+4.45*	*+2.51*
5	*3.90*	*+3.86*	*+3.03*	*+3.91*	*3.52*	*+2.87*	*+4.88*	*+2.94*
6	*7.20*	*+0.21*	*+0.94*	*+10.86*	*12.19*	*+2.54*	*12.19*	*+2.74*
7	*7.01*	*+1.09*	*+1.00*	*+6.29*	*9.44*	*+2.17*	*+2.51*	*+2.46*
8	*+1.21*	*+0.34*	*7.03*	*+7.91*	*0.22*	*+0.44*	*+2.91*	*9.04*
9	*6.53*	*+0.68*	*+0.50*	*+8.41*	*8.75*	*+0.73*	*+3.2*	*+0.29*
10	*7.47*	*+2.80*	*+5.66*	*+20.43*	*8.66*	*+2.4*	*+4.1*	*+17.69*
11	*5.79*	*+3.86*	*+2.71*	*+7.59*	*7.20*	*+1.18*	*+5.56*	*+3.32*
12	*6.62*	*+1.83*	*+1.88*	*+6.76*	*7.37*	*+1.01*	*+5.39*	*+3.15*
13	*5.58*	*+3.37*	*+2.19*	*+4.41*	*5.24*	*+1.56*	*+3.16*	*+7.39*
14	*4.89*	*+3.68*	*+2.88*	*+5.10*	*5.94*	*+0.86*	*+2.46*	*+6.69*
15	*+0.65*	*+1.07*	*6.93*	*+0.88*	*4.95*	*+1.44*	*+3.45*	*+1.51*
16	*9.20(T)*	*+1.02*	*9.20(T)*	*9.20(T)*	*13.63*	*13.63(T)*	*13.63(T)*	*13.63(T)*
17	*+0.01*	*+0.22*	*8.14*	*+9.92*	*13.28*	*+1.45*	*13.28(T)*	*+1.65*
18	*+0.9*	*7.41*	*+0.73*	*+10.65*	*13.09*	*+1.64*	*13.09(T)*	*+1.64*
19	*+0.8*	*+0.04*	*8.01*	*+5.29*	*10.16*	*+1.45*	*+1.79*	*+1.74*
20	*+0.03*	*+0.09*	*8.01*	*+5.29*	*10.23*	*+1.38*	*+1.72*	*+1.67*
21	*+2.45*	*+0.34*	*7.03*	*+7.91*	*0.7*	*+0.44*	*+2.91*	*9.04*
22	*7.30*	*+6.78*	*+5.83*	*+20.60*	*8.89*	*+2.17*	*+3.87*	*+17.46*
23	*8.13*	*+2.14*	*+5.00*	*+19.77*	*9.41*	*+1.65*	*+3.35*	*+16.94*
24	*7.95*	*+1.70*	*+0.55*	*+5.43*	*8.11*	*+0.27*	*+4.65*	*+2.41*
25	*7.42*	*+1.53*	*+0.35*	*+2.57*	*6.31*	*+0.49*	*+2.09*	*+6.32*
26	*9.45*	*9.45(T)*	*9.45(T)*	*9.45(T)*	*14.51*	*14.51(T)*	*14.51(T)*	*14.51(T)*
27	*10.15*	*+0.07*	*10.15(T)*	*10.15(T)*	*14.44*	*14.44(T)*	*14.44(T)*	*14.44(T)*
28	*+2.03*	*+0.04*	*8.01*	*+5.29*	*10.81*	*+0.8*	*+1.14*	*+1.09*
29	*+1.91*	*+0.22*	*8.14*	*+9.92*	*14.09*	*+0.64*	*14.09(T)*	*+0.84*
30	*8.86*	*+5.22*	*+4.27*	*+19.04*	*9.83*	*+1.23*	*+2.93*	*+16.52*

(T) indicates the temporal correction was used due to the correction failure

2.6.2.Results of Random-Error Evaluation

The KR approach and LD approach hereafter refer to the versions with the adaption mentioned in the last sub-section. The comparison between the four correction methods based on MAE and MAPE averaged over 10 realizations is presented in Figure 9.

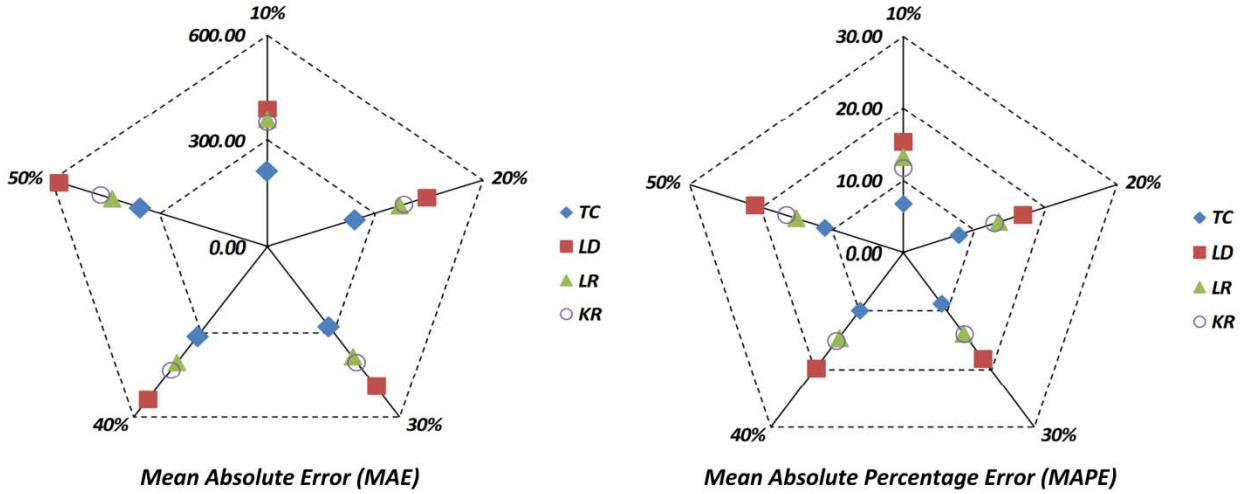


Figure 9 Comparison between Correction Methods

Clearly, the temporal correction method enjoys the best performance at all error levels. Specifically, for error level 10%, the MAPE of the temporal correction method is only 6.76% compared to 15.17% for the LD approach, 13.06% for the LR approach, and 11.58% for the KR approach. Even when the error level is 50%, the temporal correction method still achieves a MAPE of 11.08%. Considering that the percentage of errors within a dataset usually does not surpass 50%, the TC approach essentially provides very reliable correction performance for non-incident conditions (in the 2008 dataset used, only 33 days out of 314 days had error level greater than 50%).

The LR approach maintains a relatively steady performance at all error levels. The MAPEs for the five error levels range from 13.06% to 15.09% while the MAPE ranges for other three approaches each exceed 5%. This steady trend is determined by the inherent characteristics of the linear regression method. The linear relationship used for estimation essentially predicts the expected station flow conditioned on an individual lane flow. In addition, the station flow can still be estimated even if there is only one valid lane flow. Therefore, even if the error level is high, the performance of LR approach is still stable.

The performance of the lane distribution (LD) approach is not as good as anticipated. The lowest MAPE for error level 10% is higher than that reported in Smith and Conklin's (2009) original study. Their work reported that the LD approach obtained less than 8% MAPE. However, it would be misleading to say that the LD approach is not transferable to other situations. The difference between the results can be ascribed to several reasons. The original study used 10-minute aggregate data, which had less noise due to aggregation while the dataset for this study has a higher temporal resolution of one minute. Another reason is that the freeway section of this study has 5 lanes, which is different from the 3-lane situation (without exit lane) in the original study. Though only information from the three left-most lanes was used for estimating the station flow, the underlying distinct traffic pattern can lead to significantly different results. In addition, though the data for incident duration was excluded, it cannot be guaranteed that the incident impact was completely eliminated from the dataset of this study. This is because the reported incident duration may be inaccurate due to documenting errors or inaccuracies. Since being free of incidents is an explicit requirement for application of the LD approach, potential incident impact may undermine the performance of the LD approach.

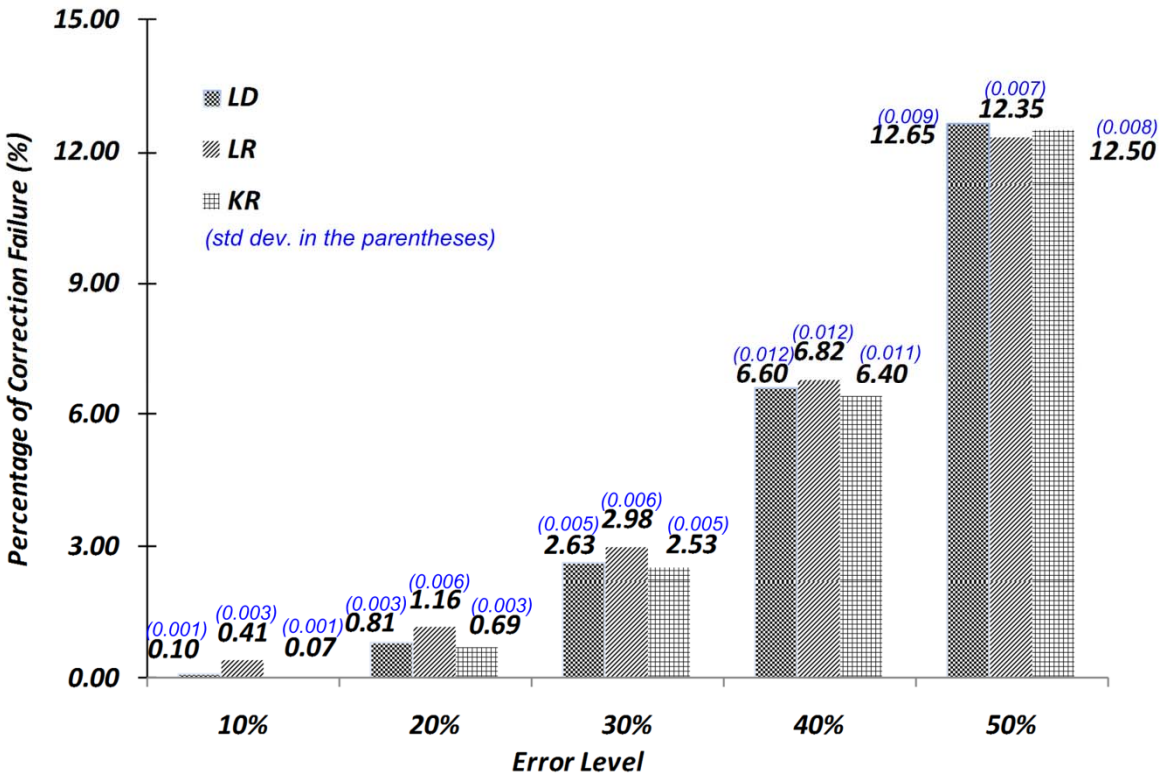


Figure 10 Percentage of Correction Failures (Averaged over 10 Realizations)

As mentioned in Section 4, the LR, KR and LD approaches require at least one functional detector in order to derive the station flow. Therefore, it is beneficial to compare the procedures in terms of their correction failure rate. Note that the correction failure indicates that the procedure failed to produce an estimate of station flow and had to use the TC estimate as a replacement. The 10-replication average and standard deviation for the percentage of correction failure are shown in Figure 10. For error levels 10% to 30%, all three correction procedures enjoy correction failure rates less than 5% and approximately 6% for the 40% error level. This suggests that they are robust under most circumstances considering that most dataset usually have less than 50% error percentage.

2.6.3.Performance of Correction Methods by Time of Day

To examine time of day effects, the MAPEs were calculated for each hour of the target day, April 22, 2008. The hourly MAPEs are averaged across the 10 random realizations. The results are visualized through Figure 11 to Figure 15. Recall that one *detector* malfunction (or more) leads to an unreliable station flow for a given time, which calls for correction. The maximum possible *station* corrections made for an hour is 60. (Recall also that the random errors are distributed among the possible individual detector readings).

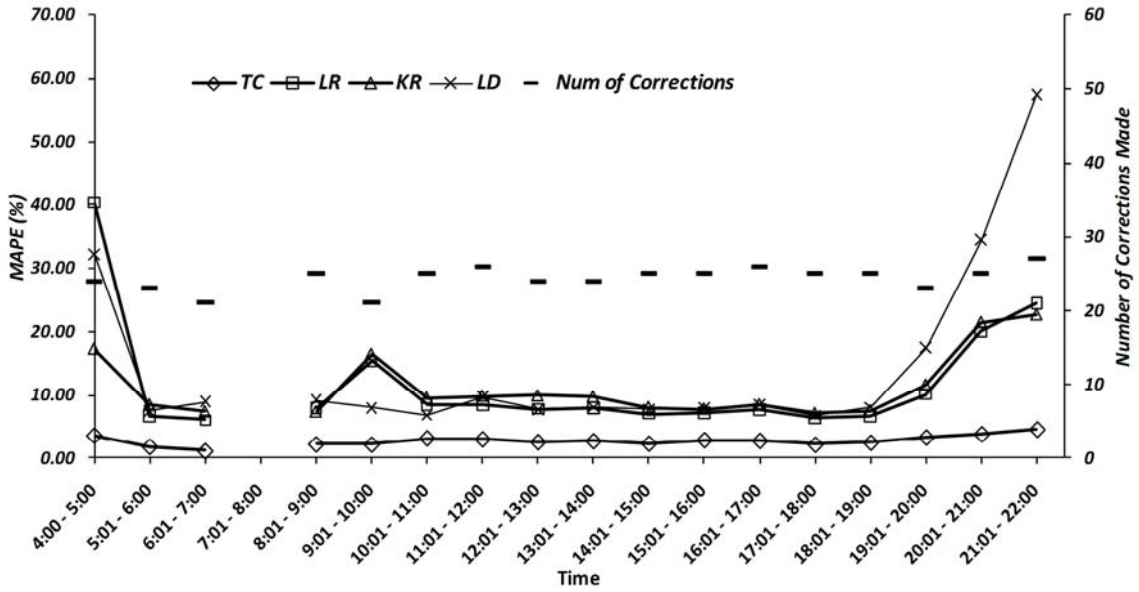


Figure 11 MAPE by Time of Day for Error Level 10%

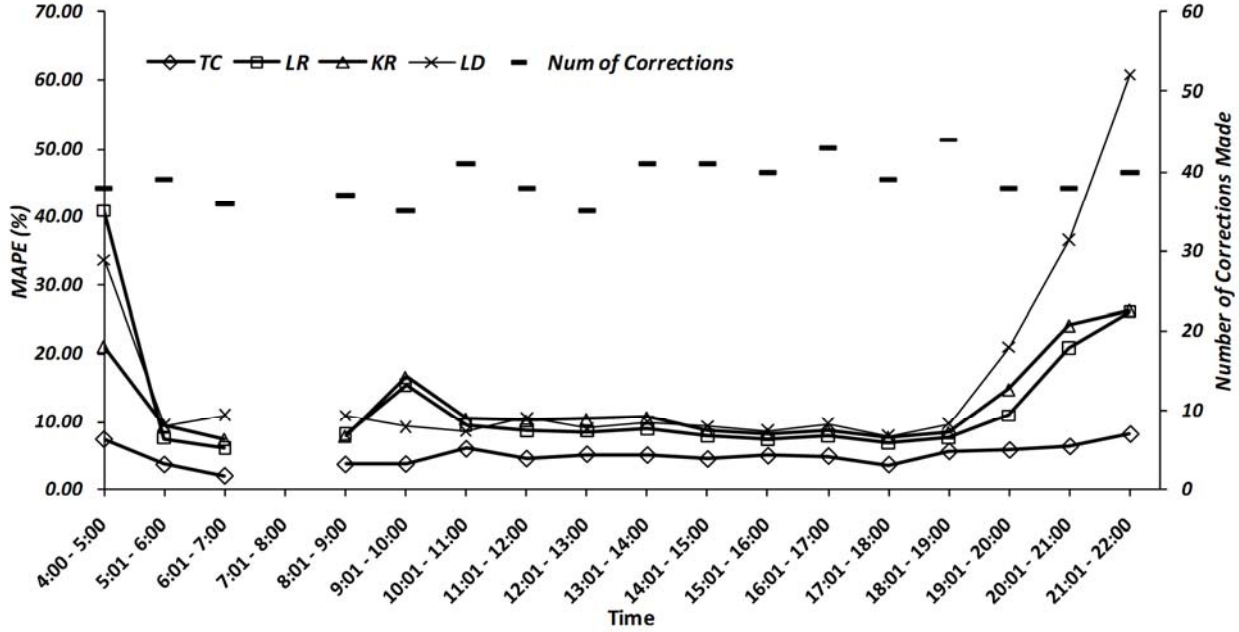


Figure 12 MAPE by Time of Day for Error Level 20%

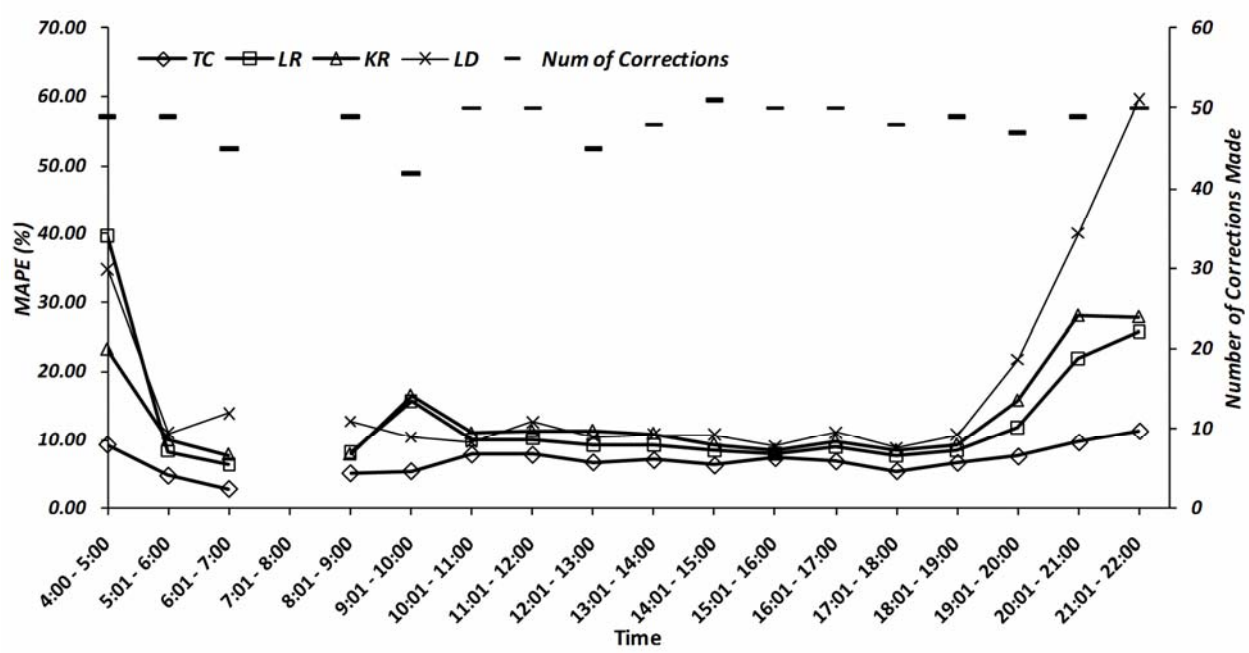


Figure 13 MAPE by Time of Day for Error Level 30%

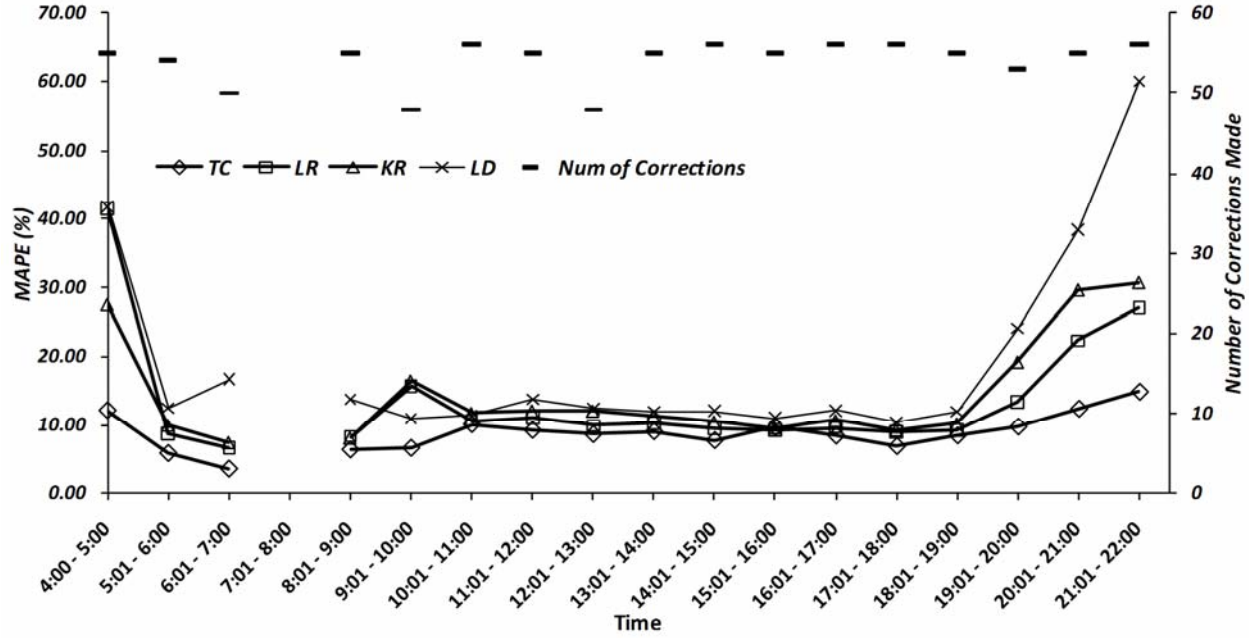


Figure 14 MAPE by Time of Day for Error Level 40%

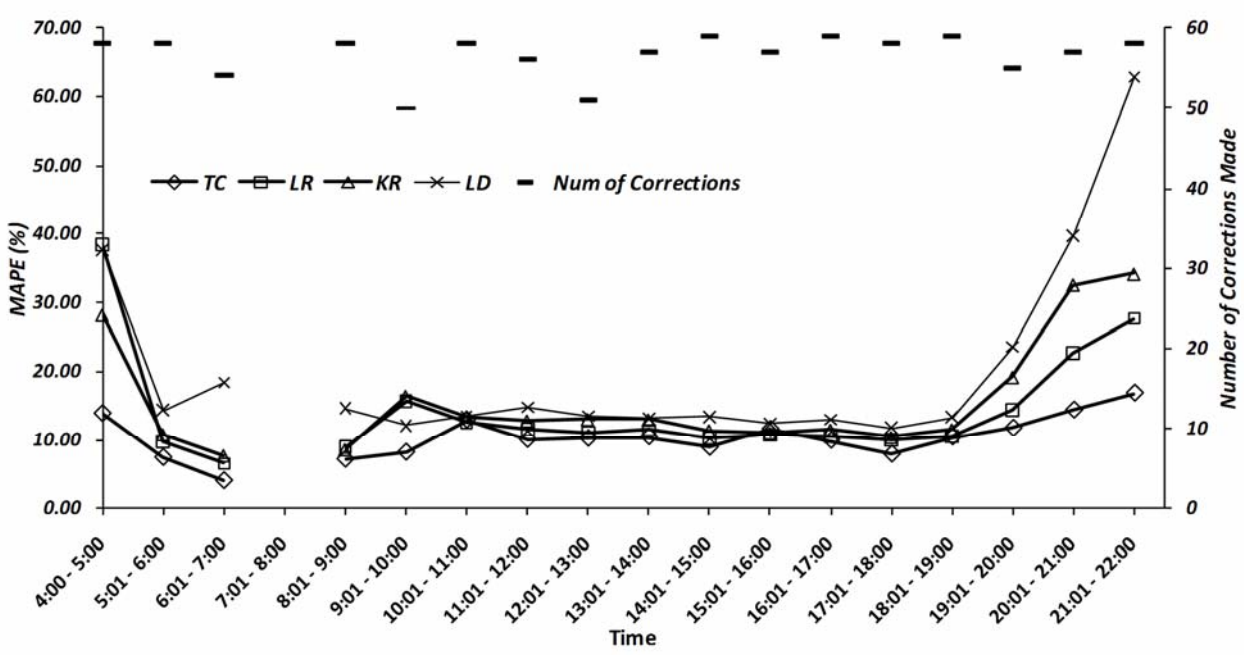


Figure 15 MAPE by Time of Day for Error Level 50%

The gap for 7:01 to 8:00 is because the original dataset does not have correctly reported flows. Therefore, it is meaningless to calculate any performance measurements due to lack of real values. It can be seen that the number of corrections made for each hour is relatively steady. Hence, the temporal variation of performance is due to the detector malfunction combinations and possibly traffic flow variations instead of the error quantity.

At all error levels, the temporal correction generally outperformed the other three procedures. The exceptions are time slot 15:00 to 16:00 for error levels 40% and 50%. In both scenarios, the LR approach, followed by the KR approach, achieved the lowest MAPE. This indicates that the LR and KR approaches outperformed the TC approach under certain circumstances. One example is the time 15:35:00 at which the real station flow is 2280 veh/hr and the working detectors for this time slot under one realization of the 40% error level are detectors 605, 607 and 609. The estimate by using the TC approach is 2979 veh/hr. This estimate was derived from the individual lane flows for detectors 601 and 603, which are 675 veh/hr and 1224 veh/hr while the observed values are 420 veh/hr and 780 veh/hr, respectively. The predicted values for the lane flows of detectors 601 and 603 overestimate the readings. A possible reason is that the temporal correction method exploits the past traffic pattern. The flow for this particular time slot deviates from the average of the flows at the same time in the past, which leads to an inaccurate final estimate. The estimate by the LR approach is 2411veh/hr and 2255 veh/hr by the KR approach (in this realization), which are very good estimates. As the general trend suggests, the temporal correction still has the best performance in terms of estimation accuracy though cases of lower accuracy occurred when the flow of a particular time slot deviated from the historical value.

The performance of the LR, KR and LD approaches exhibited relatively high MAPE for the time slot of 4:00 to 5:00 and deteriorated for the duration after 19:00. Selected individual data points between these time durations at the 10% error level are more closely examined. The duration of 4:00 to 5:00 in the morning exhibits very light traffic conditions. For example, the real lane flows at 4:10 AM were 0(d601), 0(d603), 300(d605), 0(d607) and 0(d609). In one realization, the malfunctioned detector was detector 603. Therefore, the final station flow was derived based on at least one zero flow reading (detector 601). The estimated station flow was 1221 veh/hr using the LR approach, which is four times the real station flow. Similar cases were found for the KR and LD approaches for the time duration after 20:00. These

findings suggest that the LR, KR and LD approaches may produce inaccurate results under light conditions especially when the estimates are based on zero flow readings.

2.7. Conclusions

In this study, four spatial correction procedures were examined for non-incident related detector data. The first approach, temporal correction exploited the inherent temporal trend of historical traffic. The spatial correction based on linear regression (LR), a proposed modification of a previous approach, uses the relationship between the individual detector flow and station flow. The third approach proposed in this study is also a spatial correction method. A unique feature of the proposed spatial correction procedure was incorporation of lane use percentage into the correction process through kernel regression (KR). As a comparison benchmark, the correction method based on lane distribution (LD) developed by previous researchers was included as the fourth method.

To comprehensively compare the correction procedures, both systematic evaluation and random-error evaluation were conducted. After the results of systematic evaluation were analyzed, it was found that adaption was needed for the KR and LD approaches. Specifically, the individual lane flows provided by the detectors on particular general purpose lanes were discovered to produce the more accurate estimates. The two correction procedures (kernel regression and lane distribution) were revised in light of this finding and their station flow estimates were compared to those of the temporal correction and the LR approach at five error levels, which was considered as the random-error evaluation.

The results of the study are for a specific detector station that has five lanes, one of which is an exit lane. This configuration is different from numerous previous studies and may have a significant impact on the performance of the spatial correction procedures. The transferability of the results to other facilities, lane quantities and configurations, and distance from ramps is an area for future study. However, several practical recommendations can still be made.

2.7.1.Summary of Findings and Practical Recommendations

First, it is critical to calibrate the correction methods according to location-specific characteristics since each detector has different significance for the correction accuracy. This is especially warranted for approaches involving lane distribution. Information from more individual detectors does not necessarily lead to more accurate performance as shown by the results of the two systematic evaluations. Selection rules as to which individual lane flows to include need to be constructed before applying the kernel regression (KR) and lane distribution (LD) approaches. In this study, including the information from the exit and adjacent lanes may undermine the accuracy of these two approaches.

Second, the systematic evaluation indicated that the temporal correction generally produced the most accurate results for most error configurations. However, as the number of malfunctioned detectors increased, the KR approach provided better results under certain error combinations. In addition, the linear regression (LR) approach possesses robustness, which can achieve accuracy very close to that of the best approaches under various error configurations and thus it could be a viable choice for a dataset with diverse error configurations.

It is beneficial to know the distribution of error configurations and select the correction procedure accordingly. If one error configuration is dominant in the target dataset, the best approach corresponding to this specific error configuration should be chosen. For example, it is not uncommon that a detector malfunctions for an extended period of time while its –in-station counterparts work properly. The suitable correction method for this particular time span can be selected according to the error configuration to achieve the best performance.

Third, the time of day performance assessment confirmed the superior performance of the temporal correction approach. Overall, this approach outperformed the others at all random error levels though sporadic cases in which other methods were better did exist. Associated with time of day are traffic conditions, which were important to the spatial correction procedures, especially when valid observations of zero flow were incorporated into an average. Thus, time of day and traffic conditions should be considered when selecting a correction procedure.

In addition to performance-based suggestions, considerations should be given to practical implementation. Though the temporal correction method generally provides the most accurate results, it requires the archival of the historical traffic data at all times. One query to the historical data is necessary for each individual lane correction. This may be time-consuming for certain applications in which efficiency is greatly desired such as real-time correction of traffic data streams for multiple locations. The storage space and processing time might become prohibitively expensive. When the size of the target dataset is small or speed is not the top priority, the temporal correction would be an ideal pick out of the four procedures examined, provided that the dataset is incident free. Therefore, it is important to take into the characteristics and requirements of the application into consideration when choosing the correction method. When the accuracy is not strictly demanded, correction speed is an issue, or a diverse set of error configurations are present, the spatial correction based on linear regression can be used.

CHAPTER 3 AUTOREGRESSIVE INTEGRATED MOVING AVERAGE MODEL FOR SHORT-TERM FREEWAY TRAFFIC FLOW FORECASTING

To provide input to origin-destination (OD) matrix prediction, the traffic flows are forecast using the Auto Regressive Integrated Moving Average (ARIMA) model on a short-term basis. This chapter starts with an introductory discussion of modeling traffic volume series. Then it reviews the relevant previous studies. The framework of Box-Jenkins ARIMA modeling is then presented. The modeling process for Interstate 66 and results are discussed in the fourth section. The final section draws some conclusions.

3.1. Introduction

In addition to the application of OD prediction, short-term traffic forecasting is essential for various applications such as tactical decisions and traffic system operations, advanced warning in monitoring systems, and implementation and control of long-range plans (Nihan and Holmesland, 1980). Furthermore, accurate forecasts are important for implementation and evaluation of real-time control strategies in a feedback system, such as within the demand-responsive ramp-metering control environments (Lee and Fambro, 1999) and incident detection (Ahmed and Cook, 1980), because traffic flow is one of the fundamental types of data for traffic control and the general planning process. Within the context of advanced traffic management system, the role of accurate and reliable traffic flow data is crucial to satisfy the objectives of intelligent transportation systems (Lee and Fambro, 1999). To improve the responsiveness and satiate the sophistication of the transportation systems, forecasts of future status of the transportation system, usually characterized by traffic flow, density and speed, are in demand.

Currently, individual travelers frequently face the challenge of making their trip decisions based on current, rather than short-term, anticipated traffic conditions. The dynamic nature of traffic conditions necessitates the use of short-term forecasting tools to provide transportation users with the expected travel times and delays on main routes in short time intervals, possibly in the range of 5, 10, or 15 min time frames (Ishak and Al-Deek, 2002). Short-term predicted travel times and delays may have a significant impact on the travelers' pre-trip and en route decisions in terms of route selection, departure time, and possibly mode selection, especially at areas where public transit is competitively available (Ishak and Al-Deek, 2002). Essentially, efficient dissemination of predicted travel information should help spread congestion over time and space. To achieve this objective, short-term traffic prediction models are important in the sense that they characterize the dynamic nature of traffic conditions which then help the travelers at the onset of the trip and even en route.

3.2. Previous Studies for Traffic Forecasting

Due to the significance of traffic forecasting as previously discussed, various techniques other than ARIMA modeling have been used in the area of traffic forecasting. The exponential filtering technique for traffic flow forecasting was employed by Ross (1982) and later expanded to the general digital filtering technique by Coifman (1996). The nonparametric forecasting method was applied using the k-nearest-neighbor approach to forecast short-term freeway traffic (Davis and Nihan, 1991; Tao et al., 2010). The Kalman filtering method was used for traffic volume forecasting both from the perspectives of traffic time-series (Okutani and Stephanedes, 1984) and traffic network level (Whittaker et al., 1997). Traffic forecasting has also been studied by the use of different neural networks with different configurations (Taylor and Meldrum, 1995; Cheu, 1998; Ruimin and Huapu, 2009).

The time-series technique has been widely used in the area of traffic forecasting. Ahmed and Cook (1979; 1980) were among the early attempts to use the Box and Jenkins method to forecast freeway traffic volume. They fitted an ARIMA (0,1,3) model to the traffic volume series and applied the model for freeway incident detection. They compared the short-term traffic volume forecasts of the ARIMA (0, 1, 3) model to those obtained by double exponential smoothing, simple moving average (with orders of 5, 10,

and 20), and exponential smoothing with adaptive response (the value of α is changed for each one-step forecast). The four models were developed and evaluated on 166 data sets obtained from three surveillance systems: Los Angeles (20-sec intervals), Minneapolis (30-sec intervals), and Detroit (60-sec intervals). The authors found that the ARIMA (0, 1, 3) forecasts were the most accurate in terms of the mean absolute error and mean square error. Levin and Tsao (1980) applied the ARIMA methodology to both volume and occupancy series, and they compared two models to select a proper model for the use of freeway traffic volume forecasting. They also compared the performance of the ARIMA (0, 1, 1) model with the Illinois Department of Transportation Traffic Systems Center's ARIMA (0, 1, 0) model [actually an ARIMA (3,1,0) with φ1 =φ2 =φ3 = 1/3]. The two models were tested on traffic data from the Dan Ryan Expressway in Chicago. The ARIMA (0, 1, 1) model was superior. Different from previous works, Dhingra et al. (1993) focused on the application of four ARIMA models for forecasting truck traffic volume.

Some researchers also endeavored in incorporating other techniques into the ARIMA modeling framework. An application of the classification motive to the ARIMA model was conducted with Kohonen self-organizing maps (Van Der Voort et al., 1996). In this study, an ARIMA(p, 0, q) with p = 2 or 3 and q =1 or 2 was applied to clustered outputs from a 15-row by 20-node hexagonal Kohonen map. A comparison by Smith (1995) of neural network, nearest-neighbor, historical average, and ARIMA models showed that an ARIMA (2, 1, 0) model did better than the historical average model and worse than the nearest-neighbor and neural network models. However, the ARIMA model was applied to only one data set and tested for only 2 days of the data set because of embedded missing values. This comparison was later repeated by Kirby et al. (1997) using 30-min traffic volume data. Each study concluded that one out of the two methods was better than the other, indicating that the performance of the models was subject to data and model structures. In the work by Szeto et al. (2009), the Seasonal ARIMA (SARIMA) model served as the demand generator for the Cell Transmission Model to simulate a urban transportation system in Ireland.

3.3. The Theory of ARIMA Modeling

This section presents the ARMA modeling framework in detail. First, the formulations of the various models and their properties are discussed. In addition, the Box-Jenkins time-series modeling philosophy is presented.

3.3.1. Weak Stationarity and Ergodicity

Stationarity

To discuss stationarity and ergodicity, the auto-covariance needs to be defined first. Given a particular realization such as $\{y_t\}_{t=-\infty}^{+\infty}$ on a time series process, a vector $\mathbf{x}_t$ which consists of the most recent (j+1) observations on y as of date t can be constructed:

$$\mathbf{x}_t = \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-j} \end{bmatrix}$$

From the joint distribution of $(Y_t, Y_{t-1}, \ldots, Y_{t-j})$ which is the underlying process generating this particular realization, the jth auto-covariance of Y_t denoted by g_{jt} is calculated as:

$$g_{jt} = E\left[(Y_t - m_t)(Y_{t-j} - m_{t-j})\right], \tag{3.1}$$

where m_t is the expectation of Y_t, i.e. $E(Y_t) = m_t$. This definition describes the covariance of Y_t with its own lagged value. The variance-covariance matrix of the random vector $\mathbf{x}_t$ is given as

$$\operatorname{cov}(\mathbf{x}_t) = \begin{bmatrix} \gamma_{0t} & \gamma_{1t} & \cdots & \gamma_{jt} \\ \gamma_{(-1)(t-1)} & \gamma_{0(t-1)} & \cdots & \gamma_{(j-1)(t-1)} \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{(-j)(t-j)} & \gamma_{(-j+1)(t-j)} & \cdots & \gamma_{(0)(t-j)} \end{bmatrix} \tag{3.2}$$

Therefore, the auto-covariance γ_{jt} is the $(1, j+1)$th element of the variance-covariance matrix in eq.(3.2).

If neither the mean μ_t nor the auto-covariance γ_{jt} depend on the date *t*, then the process for Y_t is said to be covariance-stationary or weakly stationary. Mathematically, these two requirements state as:

$$\begin{cases} E(Y_t) = \mu, \ \forall t \\ E[(Y_t - \mu)(Y_{t-j} - \mu)] = \gamma_j, \ \forall t, j \end{cases} \tag{3.3}$$

The weak stationarity implies that the covariance between Y_t and Y_{t-j} only depends on *j*, the length of time separating the two observations, and not on *t*, the date of the observation. In addition, γ_j and γ_{-j} are equal. To see this, recall the definition of auto-covariance in eq.(3.1),

$$\gamma_j = E[(Y_t - \mu)(Y_{t-j} - \mu)] \tag{3.4}$$

According to the definition of weak stationary process in eq.(3.3), the magnitude γ_j is identical for any *t*. Hence, the *t* in eq.(3.4) can be replaced by $(t+j)$, which generates,

$$\gamma_j = E[(Y_{(t+j)} - \mu)(Y_{(t+j)-j} - \mu)] = E[(Y_{(t+j)} - \mu)(Y_t - \mu)] \tag{3.5}$$

However, by the definition of auto-covariance, this expression defines γ_{-j}. Therefore, $\gamma_j = \gamma_{-j}$ follows. Equivalently, the variance-covariance matrix of $\mathbf{x}_t$ is now symmetric.

Ergoticity

Suppose there are *N* sequences $\{y_t^{(1)}\}_{t=-\infty}^{+\infty}$, $\{y_t^{(2)}\}_{t=-\infty}^{+\infty}$,…, $\{y_t^{(N)}\}_{t=-\infty}^{+\infty}$. If the observation with index *t* from each sequence is chosen, one obtains

$$\{y_t^{(1)}, y_t^{(2)}, \ldots, y_t^{(I)}\}$$

This amounts to a sample of *N* realizations of the random variable Y_t. If the probability density associated with this random variable Y_t is $f_{Y_t}(y_t)$, which is the unconditional density of Y_t, the expectation of Y_t is then given by

$$E(Y_t) = \int_{-\infty}^{+\infty} y_t f_{Y_t}(y_t) dy_t \tag{3.6}$$

This unconditional expectation of Y_t can also be viewed as the probability limit of the so-called ensemble average:

$$E(Y_t) = \operatorname{plim}_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} Y_t^{(i)} \tag{3.7}$$

Usually, what is available is a single realization of size T from the process, for example the first realization sequence denoted as $\{y_1^{(1)}, y_2^{(1)}, \ldots, y_T^{(1)}\}$. The sample mean of this realization is not ensemble average but the time average, which is expressed as:

$$\bar{y} = \frac{1}{T} \sum_{t=1}^{T} y_t^{(1)} \tag{3.8}$$

A covariance-stationary process is said to be ergodic for the mean if eq.(3.8) converges in probability to $E(Y_t)$ in eq.(3.7). Indeed, if the auto-covariance for a weakly stationary process satisfy

$$\sum_{j=0}^{\infty} |g_j| < \infty \quad , \tag{3.9}$$

then the process is ergodic for the mean.

3.3.2.The White-Noise Process

The basic building block for all ARMA processes is the white-noise process $\{\varepsilon_t\}_{-\infty}^{+\infty}$ whose elements have zero mean and finite variance σ^2:

$$\begin{cases} E(e_t) = 0 \\ E(e_t^2) = s^2 \end{cases}, \tag{3.10}$$

and the ε's are uncorrelated across time, i.e.

$$E(e_t e_t) = 0 \tag{3.11}$$

If the white-noise process satisfying eqs.(3.10) and (3.11) follows Gaussian distribution,

$$e \sim N(0, s^2), \tag{3.12}$$

the process is a Gaussian white-noise process.

3.3.3.The qth Order Moving Average Process – MA(q)

A qth-order moving average process, commonly abbreviated as MA(q), is characterized by

$$Y_t = m + e_t + q_1 e_{t-1} + q_2 e_{t-2} + \mathrm{L} + q_q e_{t-q}, \tag{3.13}$$

where $\{e_t, e_{t-1}, \mathrm{K}, e_{t-q}\}$ satisfies the requirements of white-noise series, the coefficients a's and m are deterministic real scalars. The mean of Y_t is given by taking expectation on both sides of eq.(3.13):

$$E(Y_t) = m + \sum_{j=0}^{q} q_j E(e_{t-j}) = m \tag{3.14}$$

where q_0 is defined to be unity. The variance of Y_t is computed as

$$g_0 = E\left[(Y_t - m)^2\right] = E\left[e_t + q_1 e_{t-1} + q_2 e_{t-2} + \mathrm{L} + q_q e_{t-q}\right]^2. \tag{3.15}$$

Recall that the ε's are uncorrelated across time in eq.(3.11). This implies that the expectations of the cross-terms are all equal to zero. Therefore, the variance of Y_t in eq.(3.15) simplifies to eq.(3.16).

$$g_0 = E\left[e_t + q_1 e_{t-1} + q_2 e_{t-2} + \mathrm{L} + q_q e_{t-q}\right]^2 = (1 + q_1^2 + q_2^2 + \mathrm{L}\ q_q^2)s^2 \tag{3.16}$$

The auto-covariance $g_j \ (j = 1, 2, \mathrm{K}\ q)$ is defined as:

$$\begin{aligned} g_j &= E[(e_t + q_1 e_{t-1} + q_2 e_{t-2} + \mathrm{L} + q_q e_{t-q}) \\ &\quad \times (e_{t-j} + q_1 e_{t-j-1} + q_2 e_{t-j-2} + \mathrm{L} + q_q e_{t-j-q})] \\ &= E[q_j e_{t-j}^2 + q_{j+1} q_1 e_{t-j-1}^2 + q_{j+2} q_2 e_{t-j-2}^2 + \mathrm{L} + q_q q_{q-j} e_{t-q}^2] \end{aligned} \tag{3.17}$$

For $j > q$, there are no e's with identical dates in the definition of g_j and thus the expectation is zero. To summarize, the auto-covariance of an MA(q) process is shown in

$$g_j = \begin{cases} (1 + q_1^2 + q_2^2 + \mathrm{L}\ q_q^2)s^2, & j = 0 \\ (q_j + q_{j+1} q_1 + q_{j+2} q_2 + \mathrm{L} + q_q q_{q-j})s^2, & j = 1, 2, \mathrm{K}\ q \\ 0, & j > q \end{cases} \tag{3.18}$$

As an illustrative example, for an MA(2) process defined as:

$$Y_t = m + e_t + q_1 e_{t-1} + q_2 e_{t-2} \tag{3.19}$$

The auto-covariance is computed using eq.(3.18) and results are

$$\begin{cases} g_0 = \left(1 + q_1^2 + q_2^2\right)s^2 \\ g_1 = \left(q_1 + q_2 q_1\right)s^2 \\ g_2 = q_2 s^2 \\ g_3 = g_4 = \text{L} = 0 \end{cases} \tag{3.20}$$

The MA(q) model in eq.(3.13) can be rewritten in a more compact form as in eq.(3.21) with $q_0 = 1$:

$$Y_t = m + e_t + q_1 e_{t-1} + q_2 e_{t-2} + \text{L} + q_q e_{t-q} = m + \sum_{j=0}^{q} q_j e_{t-j} \tag{3.21}$$

Consider the process when q approaches infinity,

$$Y_t = m + \sum_{j=0}^{q} y_j e_{t-j} = m + y_0 e_t + y_1 e_{t-1} + y_2 e_{t-2} + \text{L} \tag{3.22}$$

This process defined in eq.(3.22) can be described as an MA(∞) process. Note the notational difference for the coefficients. y's are used to represent coefficients in an infinite-order MA process while a's those in a finite-order coefficients.

3.3.4.The pth-order Autoregressive Process- AR(p)

A pth-order autoregression, denoted by AR(p), satisfies eq.(3.23).

$$Y_t = c + f_1 Y_{t-1} + f_2 Y_{t-2} + \text{L} + f_p Y_{t-p} + e_t \tag{3.23}$$

Provided the roots of the characteristic equation of the AR(p) process in eq.(3.24),

$$1 - f_1 z - f_2 z^2 - \text{L} - f_p z^p = 0 \tag{3.24}$$

all lie outside the unit circle, the AR(p) process can be converted to the covariance-stationary infinite-order MA(∞) process, which has the representation as

$$Y_t = m + y(L)e_t \tag{3.25}$$

where

$$y(L) = (1 - f_1 L - f_2 L^2 - \text{L} - f_p L^p)^{-1} \tag{3.26}$$

and

$$\sum_{j=0}^{\infty} |y_j| < \infty \tag{3.27}$$

The first-moment of a stationary AR(p) process can be found by taking expectation of both sides of eq.(3.23). Recall the characteristics of the white-noise process in eq.(3.10), one writes:

$$m = c / (1 - f_1 - f_2 - \text{L} - f_p) \tag{3.28}$$

Writing the AR(p) process in its mean-deviation form as eq.(3.29),

$$Y_t - m = c + f_1(Y_{t-1} - m) + f_2(Y_{t-2} - m) + \text{L} + f_p(Y_{t-p} - m) + e_t \tag{3.29}$$

the auto-covariance are found by multiplying both sides of (3.29) by $(Y_{t-j} - m)$ and taking expectations. This leads to eq.(3.30).

$$g_j = \begin{cases} f_1 g_{j-1} + f_2 g_{j-2} + \text{L} + f_p g_{j-p}, & j \geq 1 \\ f_1 g_1 + f_2 g_2 + \text{L} + f_p g_p + s^2, & j = 0 \end{cases} \tag{3.30}$$

Since the process is weakly stationary, $g_j = g_{-j}$, the system of equations in (3.30) for $j = 0,1,\ldots,p$ can be solved for $g_0, g_1, \ldots, g_p$ as functions of $s^2, f_1, f_2, \ldots, f_p$. In fact, the vector $(g_0, g_1, \ldots, g_{p-1})'$ is given by the first p elements of the first column of the $(p^2 \times p^2)$ matrix $s^2[\mathbf{I}_{p^2} - (\mathbf{F} \otimes \mathbf{F})]^{-1}$ where $\mathbf{F}$ is defined as eq.(3.31) and $\otimes$ indicates the Kronecker product.

$$F = \begin{bmatrix} f_1 & f_2 & f_3 & \mathrm{L} & f_{p-1} & f_p \\ 1 & 0 & 0 & \mathrm{L} & 0 & 0 \\ 0 & 1 & 0 & \mathrm{L} & 0 & 0 \\ 0 & 0 & 1 & \mathrm{L} & 0 & 0 \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ 0 & 0 & 0 & \mathrm{L} & 1 & 0 \end{bmatrix} \tag{3.31}$$

3.3.5.The Auto Regressive Moving Average Process - ARMA(*p*,*q*)

An ARMA(*p*,*q*) process includes both auto regressive and moving average terms. It takes the form as eq.(3.32).

$$Y_t = c + f_1 Y_{t-1} + f_2 Y_{t-2} + \mathrm{L} + f_p Y_{t-p} + e_t + q_1 e_{t-1} + q_2 e_{t-2} + \mathrm{L} + q_q e_{t-q} \tag{3.32}$$

Or in lag operator form, it is written as eq.(3.33).

$$(1 - f_1 L - f_2 L^2 - \mathrm{L} - f_p L^p) Y_t = c + (1 + q_1 L + q_2 L + \mathrm{L} + q_q L^q) e_t \tag{3.33}$$

If the roots of the characteristic polynomial of the auto regressive terms

$$f(L) = 1 - f_1 L - f_2 L^2 - \mathrm{L} - f_p L^p \tag{3.34}$$

are outside the unit circle. Both sides of (3.33) can be divided by $f(L)$ to obtain

$$Y_t = m + y(L) e_t \tag{3.35}$$

where

$$\begin{cases} y(L) = \dfrac{1 + q_1 L + q_2 L + \mathrm{L} + q_q L^q}{1 - f_1 L - f_2 L^2 - \mathrm{L} - f_p L^p} \\ m = c / (1 - f_1 - f_2 - \mathrm{L} - f_p) \\ \sum_{j=0}^{\infty} |y_j| < \infty \end{cases} \tag{3.36}$$

3.3.6.The Box-Jenkins Modeling Philosophy

The Box-Jenkins modeling philosophy adopts a three-step iterative scheme, which consists of identification, estimation and diagnostic checking.

(1) Identification: Calculate the sample autocorrelation function (ACF) and partial autocorrelation function (PACF) for different lags. We compare the estimated ACF and PACF with various theoretical ACF's and PACF's to find a match. A tentative model is chosen when its ACF and PACF best match the estimated ACF and PACF.

(2) Estimation. The tentative model identified is fitted to the data to get estimates of its parameters. These coefficients are examined for stationarity, invertibility, statistical significance and other quality measures.

(3) Diagnostic-checking. At this stage, the hypothesis that the residuals of the estimated model are white-noise is checked. If they are not, one returns to the identification stage to select another model. This iteration continuous until the diagnostic-checking procedure suggests the model meets the requirement.

3.4. Modeling the Traffic Flow Series for Interstate 66

The traffic flow series used to demonstrate the Box-Jenkins time-series technique is from detector station 27, which is a mainstream detector station on Interstate 66. The data is in 5-minute resolution, which was converted from one-minute series after temporal correction discussed in Chapter 2.

3.4.1.Identification

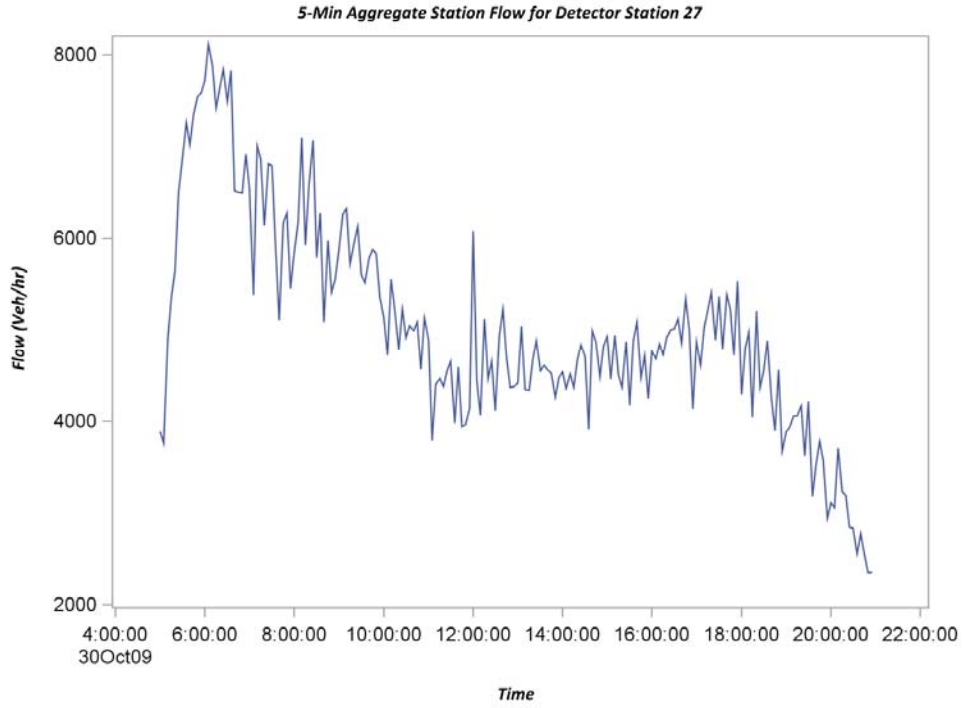


Figure 16 Series Plot of Hourly Flow for Station 27

In Figure 16, the station flow series is plotted. It can be seen that the flow series has decreasing trend. Hence, it is not stationary. The first-difference is applied to make it stationary. The differenced series is shown in Figure 17.

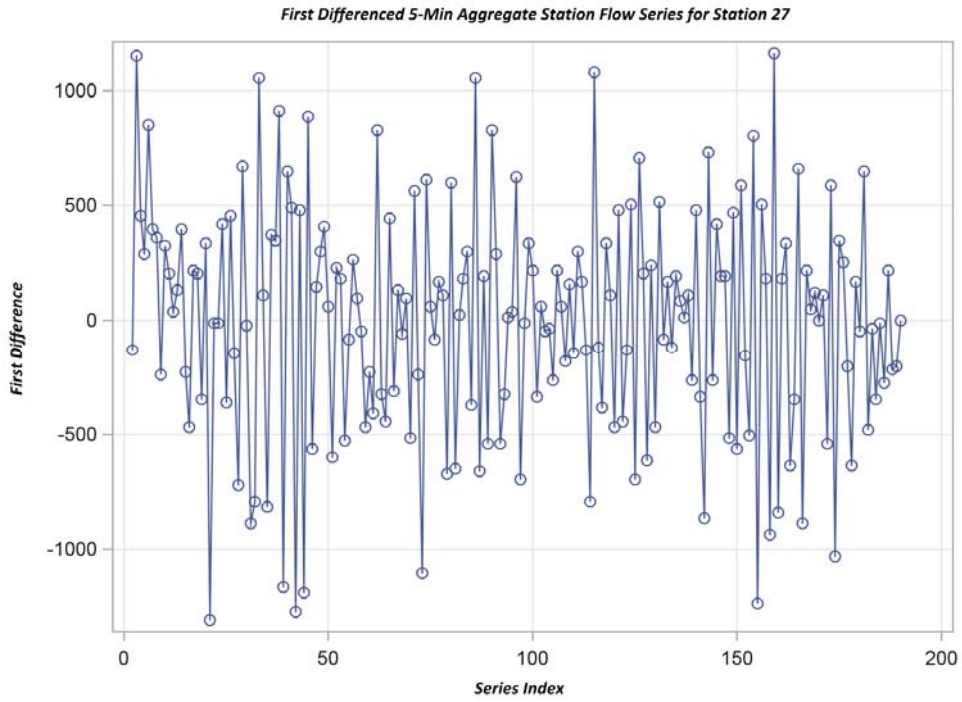


Figure 17 First-Difference Series of Hourly Station Flow

The series after first-difference is stationary since there is no obvious trend and the variance seems homogenous. Hence, we can now proceed to the calculation of the ACF's and PACF's.

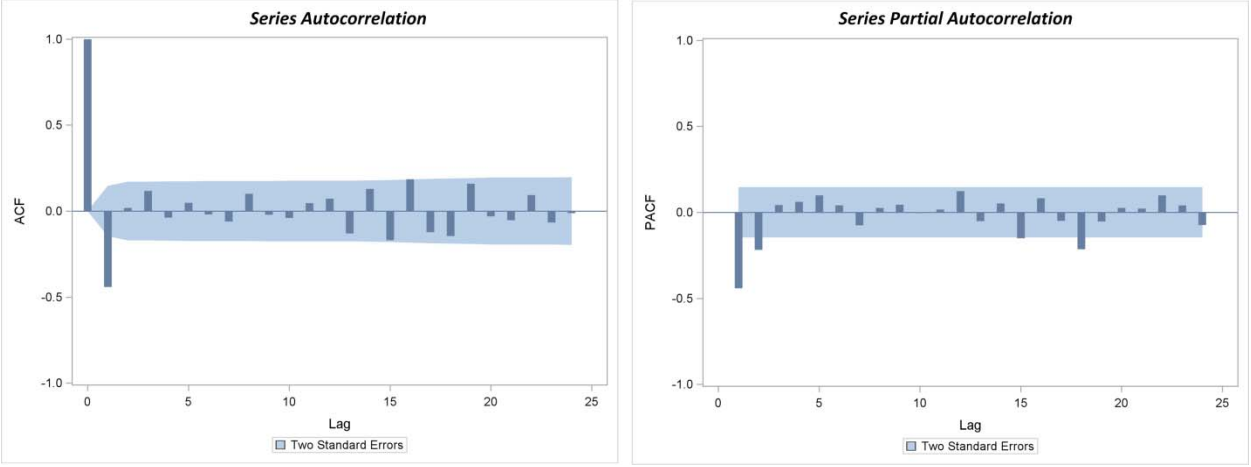


Figure 18 ACF and PACF for the First Difference Series

Figure 18 shows the ACF's and PACF's for various lags, which suggests an MA(2) model due to the cut-off ACF and dampening PACF. The parameter estimates are shown Table 2.

Table 2 Parameter Estimates for MA(2) Model

Parameter	Estimate	Std. Err.	t Value	Pr > |t|
MU	-7.8157	19.9396	-0.39	0.6955
MA1,1	0.5660	0.07214	7.85	<.0001
MA1,2	-0.1810	0.07224	-2.51	0.0131

The coefficients all enjoy statistical significance at 0.05 level. Note that the outliers have been identified and removed from the first-difference series. The diagnostic panel is shown in Figure 19.

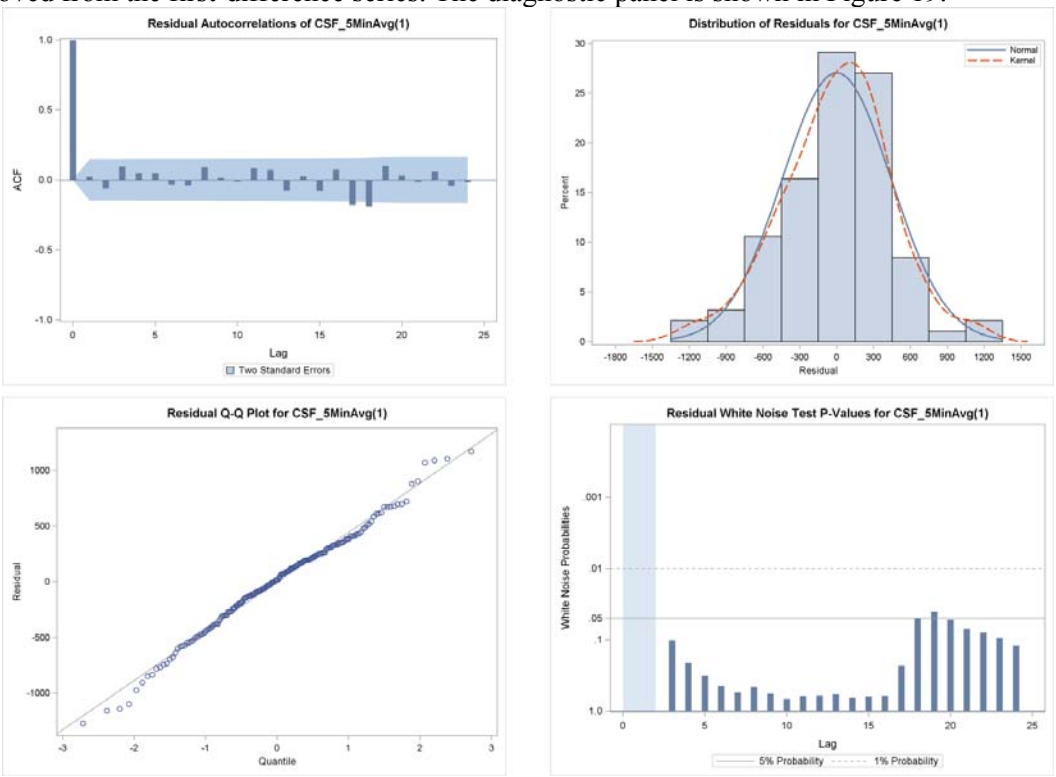


Figure 19 Diagnostic Panel for the MA(2) Model

The upper left panel is the residual autocorrelation. All lagged autocorrelation is not significantly different from zero, which supports the covariance-stationary assumption for the residual. The normal distribution of the residual is supported by the QQ plot and density plot, which suggest close alignment of Gaussianity. The White-noise plot (lower right panel) suggests that the lags until 24 are not significantly different from zero, which further confirms the validity of the white-noise residual.

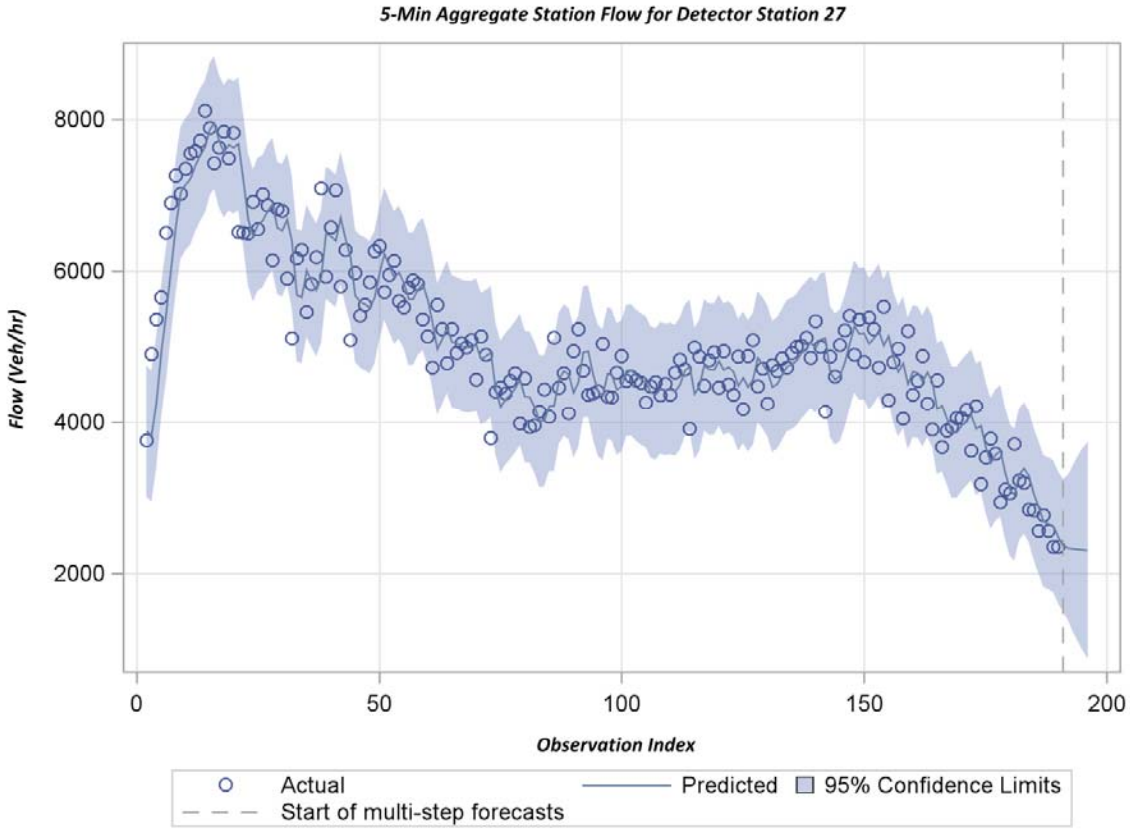


Figure 20 Forecast of the MA(2) Model

The MA(2) model can now be used for forecasting the future flows. Figure 20 shows the half-hour forecast into the future (after the dash line). The superiority of the fitness of the MA(2) model is evidenced by an MAE (Mean Absolute Error) of 344.53 and MAPE (Mean Absolute Percentage Error) of 7.07%.

3.5. Conclusions and Discussion

The station flow series was successfully modeled by an MA(2) model using the Box-Jenkins time series technique. It enjoys an MAE of 344.53 and MAPE of 7.07%. The limitation is that different models are needed for each station and the model may be different for a different time period. In other words, the model is not transferable spatially or temporally. However, once the model is fitted, it can be used to forecast the traffic hourly flow, which can subsequently be used to predict the origin-destination demands.

CHAPTER 4 DYNAMIC ORIGIN-DESTINATION MATRICES ESTIMATION

The traffic flows modeled by the Box-Jenkins time series technique are used to estimate the origin-destination (OD) matrices in a dynamic fashion. The first section discusses the QueensOD software and the second section presents the modeling process and results.

4.1. Overview of the OD Estimation Process in QueensOD

The QueensOD is an algorithm that solves the so-called synthetic OD estimation problem. The objective of the problem is to estimate the OD matrices based on the information about the magnitude of trip ends and additional links along the route of each trip (Van Aerde et al., 2003). Within the overall static synthetic OD generation problem, there are two main categories. The first exists when the routes that vehicles take through the network are known a priori (Van Aerde et al., 2003). The second arises when these routes need to be estimated concurrently while the OD is being estimated (Van Aerde et al., 2003). A prior knowledge of routes can arise automatically when there is only one feasible route between each OD pair or when observed traffic volumes are provided only for the zone connectors at the origins and destinations in the network. The first condition is common when ODs are estimated for a single intersection or arterial or for a single interchange or freeway, which is exactly our scenario for Interstate 66.

The following notations in Table 3 are used in this chapter.

Table 3 Notations for QueensOD Model

Notation	Meaning
T_{ij}	Estimated number of trips between zone *i* and *j* for the analysis period for trip purposes
T_{ij}^{k}	Estimated number of trips between zone *i* and *j* for the analysis period for trip purpose *k*
c_{ij}^{k}	Proportion of trips between zones *i* and *j* for purpose *k*. It is equal to T_{ij}^{k} / T_{ij}
P_i	Total number of trips produced by zone *i* for the analysis period
A_j	Total number of trips attracted by zone *j* for the analysis period
F_{ij}	Impedance function for travel between zones *i* and *j* during the analysis period (typically inverse function of travel time)
K_{ij}	Socioeconomic adjustment factor between zones i and j
t_{ij}	Seed trips between zones *i* and *j*
T	Total number of trips ($\sum_i \sum_j T_{ij} = \sum_i P_i = \sum_j A_j$
V_a	Actual observed link volume on link a
V_a^{c}	Volumes on link a that are closest to V_a that satisfy flow continuity.

The QueensOD model starts from the well-known gravity trip distribution formula as shown in

$$T_{ij} = P_i \left(\frac{A_j F_{ij} K_{ij}}{\sum_j A_j F_{ij} K_{ij}} \right) \tag{3.37}$$

For the trip distribution formula, the trip production and attraction rates are not measured with road counts but are estimated from land use based on trip distribution equations. QueensOD treats the trip distribution matrix as a seed matrix and then systematically uses the observed traffic flow counts in a

synthetic O-D generation process (Van Aerde et al., 2003). Because the O-D problem is underspecified, multiple O-D demands can generate identical link flows. The use of a seed matrix ensures that the optimum solution that satisfies the link flows resembles the seed solution as closely as possible (Van Aerde et al., 2003).

QueensOD employs the maximum likelihood formulation for the OD estimation problem due to the under-specified nature of the problem. The trip-based approach to defining maximum likelihood considers that the overall trip matrix is made up of uniquely identifiable individual trip makers (Van Aerde et al., 2003). The likelihood function takes the following form:

$$\max Z\left(T_{ij}, t_{ij}\right) = \frac{T!}{\tilde{O}_{ij}\left(T_{ij}!\right)} \tilde{O}_{ij} \left(\frac{t_{ij}}{\sum_{ij} t_{ij}} \right)^{T_{ij}} \tag{3.38}$$

This likelihood function is to be maximized, subject to a few constraints. The first constraint is the flow continuity, which requires the sum of all trips crossing a given link must be equal to the link flow on that link, as indicated in

$$V_a = \sum_{ij} T_{ij} p_{ij}^a, \quad \forall a \tag{3.39}$$

This simple set of equality constraints, while making the formulation complete, at times may also render the problem infeasible (Van Aerde et al., 2003). A more general formulation proposed here, therefore, is to request that the error in the link flow constraints be minimized instead of eliminated. In other words, instead of finding the most likely O-D that exactly replicates the observed link flows, the problem is reformulated as finding the most likely O-D matrix from among all those that come equally close to matching the link flows (Van Aerde et al., 2003). The mathematical expression for minimizing the error is shown in (3.40).

$$\min Z\left(T_{ij}\right) = \sum_a \left(V_a - \sum_{ij} T_{ij} p_{ij}^a \right)^2 \tag{3.40}$$

The flow continuity constraint in the form of eq.(3.40) can be integrated with the objective function in eq.(3.38) as:

$$\max \frac{T!}{\tilde{O}_{ij} T_{ij}} \tilde{O}_{ij} \left(\frac{t_{ij}}{t} \right)^{T_{ij}} - \sum_{ij} \left(l_{ij} \times 2 \left[\sum_a \left(V_a p_{ij}^a \right) - \sum_a p_{ij}^a \left(\sum_{xy} T_{xy} p_{xy}^a \right) \right] \right) \tag{3.41}$$

Taking the natural logarithm of the objective function makes the output easier to handle and allows one to use Stirling's approximation as a convenient continuous equivalent to the term ln(x!), as indicated in eq.(3.42).

$$\ln\left(T!\right) = T \ln T - T \tag{3.42}$$

The final form of the objective function of an unconstrained optimization problem is shown in

$$\max T \ln\left(\frac{T}{t}\right) - \sum_{ij} T_{ij} \ln\left(\frac{T_{ij}}{t_{ij}}\right) - \sum_{ij} \left(l_{ij} \times 2 \left[\sum_a \left(V_a p_{ij}^a \right) - \sum_a p_{ij}^a \left(\sum_{xy} T_{xy} p_{xy}^a \right) \right] \right) \tag{3.43}$$

QueensOD fully optimizes the objective function of eq.(3.43). This formulation makes use of a single approximation—namely, the Stirling's approximation—which has been shown to produce errors of less than 1% for the range of values and derivatives typically being considered in the problem (Van Aerde et al., 2003).

4.2. OD Estimation of Interstate 66 Using QueensOD

The study section of I-66 E consists of three interchanges as shown in Figure 21.

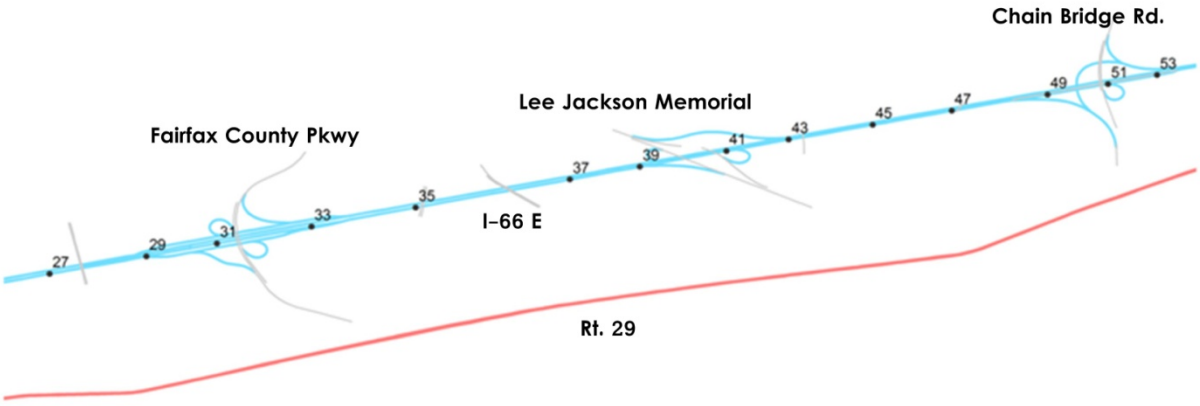


Figure 21 Study Section of I-66 E for OD Estimation

The study section is represented by an abstract network in QueensOD as shown in Figure 22.

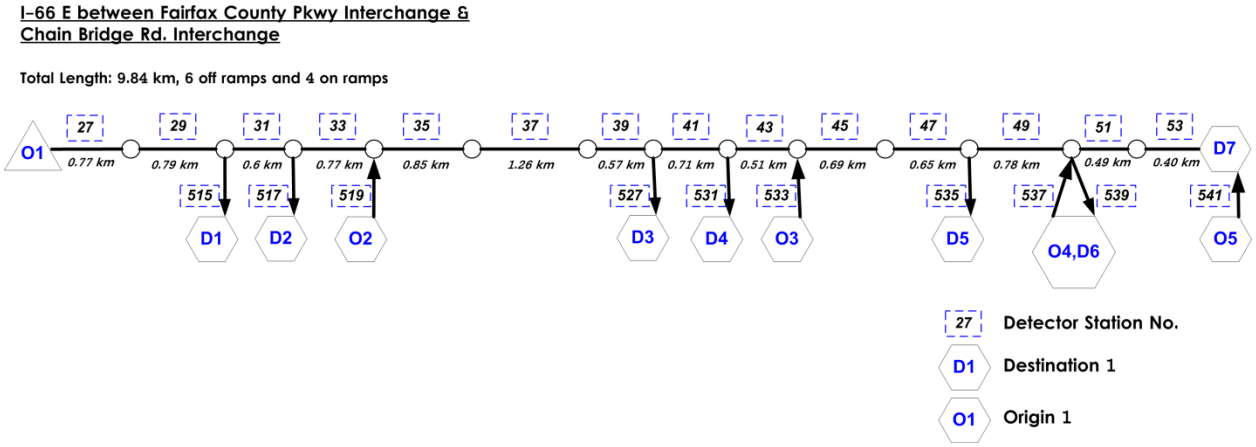


Figure 22 Study Section Represented in QueensOD

Each on-ramp is modeled as an origin and off ramp is considered as a destination. In total, there are 7 destinations and 5 origins. Note the links are numbered after the corresponding detector station number except for those whose detector station has a number larger than 500. These links are ramps and numbered as their station number minus 400. The QueensOD input files for the network are attached in the Appendix B.

The traffic flows used are from 4:00 AM to 9:00 PM on 10/30/2009. The temporal resolution is 5 minutes, which were converted from one-minute data. The one-minute data is corrected using the temporal correction approach identified as the best correction approach in Chapter 2.

The performance of the OD estimation is evaluated by checking whether the estimated flows reproduce the observed ones. The flows for part of links are organized into a panel shown in Figure 23.

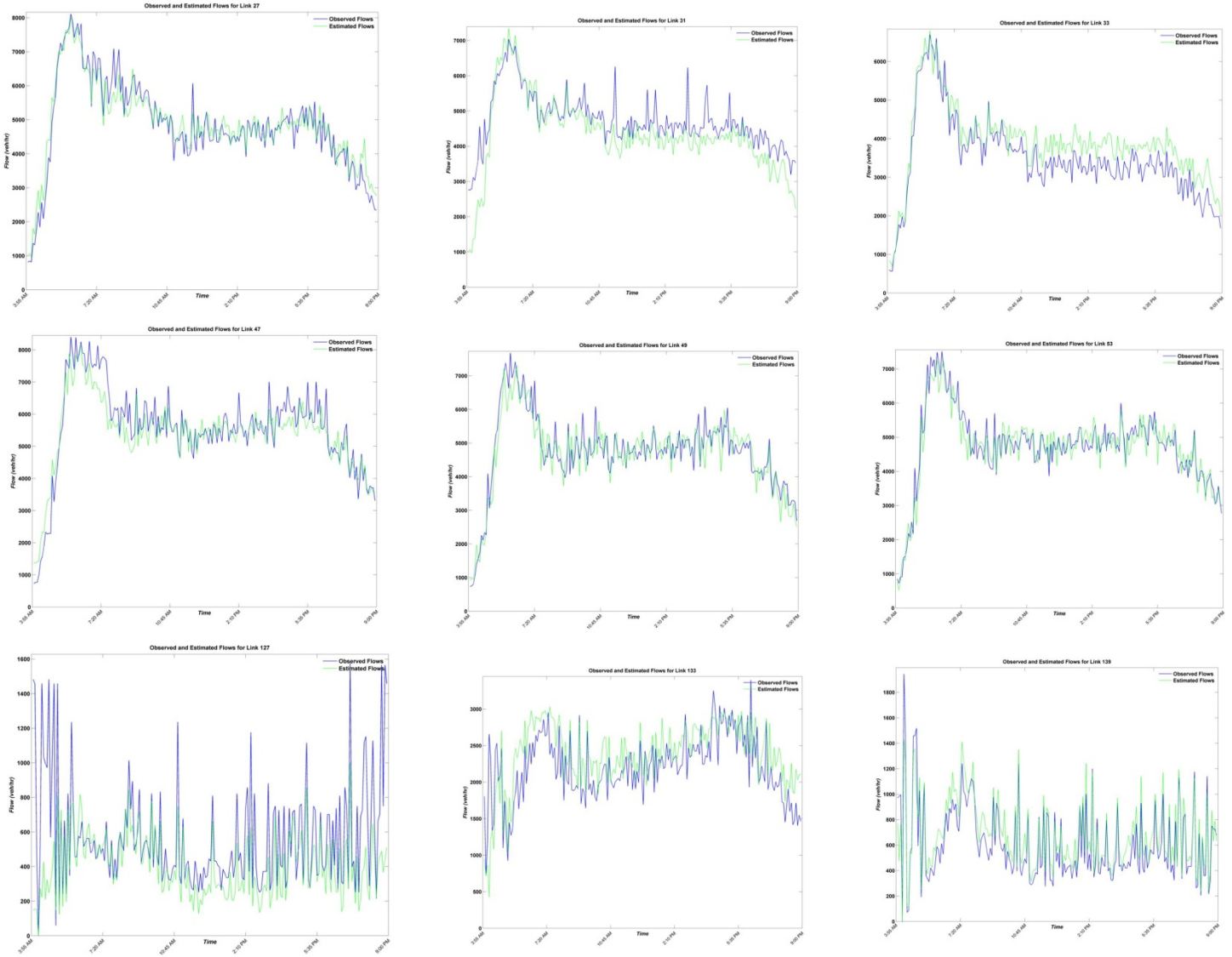


Figure 23 Comparison between Estimated Flows and Observed Flows

The numerical performance measures, namely Mean Absolute Errors (MAEs) and Mean Absolute Percentage Errors (MAPEs) are shown Table 4.

Table 4 Performance Measures Estimated Link Flows

Link No.	MAE	RMSE	MAPE
1	279.66	362.56	6.98%
2	1374.23	1414.48	42.85%
3	451.72	621.26	10.44%
4	425.01	481.26	13.96%
5	317.84	482.12	7.80%
6	295.77	455.43	7.08%
7	550.70	872.55	11.51%
8	1434.22	1810.92	28.23%
9	573.64	692.85	15.65%
10	889.94	980.17	14.73%
11	378.95	515.32	8.45%
12	332.62	440.74	7.38%
13	381.51	529.70	7.91%
14	276.49	339.96	6.31%
15	492.48	633.80	43.63%
16	232.77	272.18	33.21%
17	7.73	33.30	35.33%
18	186.28	310.89	27.72%
19	147.87	267.61	19.72%
20	294.30	375.96	14.34%
21	108.14	143.26	20.85%
22	88.07	136.02	12.89%
23	105.22	131.02	19.35%
24	86.91	231.87	3.06%

Generally, the relative errors (MAPE) are around 10% with a few higher exceptions. This is possibly due to the fact that the QueensOD requires the actual travel time as inputs when estimating dynamic OD matrices. The actual travel times were computed using the BPR function according to observed flows, which could be a source of error.

CHAPTER 5 DISCUSSIONS AND CONCLUSIONS

This project uses the Box-Jenkins time-series technique to model and forecast the traffic flows and then use the flow forecasts to predict the origin-destination matrices. First, a detailed analysis was conducted to investigate the best data correction method.

Four spatial correction procedures were examined for non-incident related detector data. The first approach, temporal correction exploited the inherent temporal trend of historical traffic. The spatial correction based on linear regression (LR), a proposed modification of a previous approach, uses the relationship between the individual detector flow and station flow. The third approach proposed in this study is also a spatial correction method. A unique feature of the proposed spatial correction procedure was incorporation of lane use percentage into the correction process through kernel regression (KR). As a comparison benchmark, the correction method based on lane distribution (LD) developed by previous researchers was included as the fourth method.

To comprehensively compare the correction procedures, both systematic evaluation and random-error evaluation were conducted. After the results of systematic evaluation were analyzed, it was found that adaption was needed for the KR and LD approaches. Specifically, the individual lane flows provided by the detectors on particular general purpose lanes were discovered to produce the more accurate estimates. The two correction procedures (kernel regression and lane distribution) were revised in light of this finding and their station flow estimates were compared to those of the temporal correction and the LR approach at five error levels, which was considered as the random-error evaluation.

After applying the temporal correction method to the data set, the station flow series was modeled using the Box-Jenkins time-series modeling framework. The station flow series was successfully modeled by an MA(2) model using the Box-Jenkins time series technique. It enjoys an MAE of 344.53 and MAPE of 7.07%. The limitation is that different models are needed for each station and the model may be different for a different time period. In other words, the model is not transferable spatially or temporally. However, once the model is fitted, it can be used to forecast the traffic hourly flow, which were subsequently used to predict the origin-destination demands.

The performance of QueensOD modeling is relatively good, which is evidenced by approximately 10% MAPE for a majority of links. Generally, the relative errors (MAPE) are around 10% with a few higher exceptions. This is possibly due to the fact that the QueensOD requires the actual travel time as inputs for dynamic OD estimation. These travel times were estimated in this study using the BPR function according to observed flows, which could be a source of error.

APPENDIX A SAS CODE FOR BOX-JENKINS MODELING

```
/*------------------------------UBJ-ARIMA Modeling-------------------------
-*/

/*ODS Graphics Settings*/
ODS GRAPHICS ON;
ODS LISTING IMAGE_DPI = 300 STYLE= STATISTICAL;

/*Convert the Sum 5-Min Data to Average 5-Min Data*/
DATA Trts.Station_27_5Min;
      SET Trts.Station_27_5Min;
      CSF_5MinAvg = CSF_5Min/5;
RUN;

/*----------STEP 1: IDENTIFICATION-----------------*/
TITLE "5-Min Data ";
PROC SGPLOT DATA = Trts.Station_27_5Min;
      SERIES X = Traffic_Time Y = CSF_5MinAvg;
RUN;
/*Findings*/
*1: The original series is not stationary.;



TITLE "5-Min 1st Diff";
PROC ARIMA DATA = Trts.Station_27_5Min
      PLOTS(UNPACK) = ALL;
      IDENTIFY
                  VAR  = CSF_5MinAvg(1)
                  NLAG = 24;
RUN;
QUIT;
/*Findings*/
*1: The first-difference series seems stationary.;
TITLE "5-Min 2nd Diff";
PROC ARIMA DATA = Trts.Station_27_5Min
      PLOTS(UNPACK) = ALL;
      IDENTIFY
                  VAR  = CSF_5MinAvg(1,1)
                  NLAG = 24;
RUN;
QUIT;
/*Findings*/
*1: The second-difference series does not present significantly different
results.;
*2: Use both difference series for estimation.
*3: The ACF cut off at 1 or 2 and PACF damps out, suggesting an MA(1 or 2)
model;

/*-----------STEP 2: ESTIMATION-------------------*/
TITLE "5-Min 1st Diff MA(1) and MA(2)";
PROC ARIMA DATA = Trts.Station_27_5Min
      PLOTS(UNPACK) = ALL;
      IDENTIFY
                  VAR  = CSF_5MinAvg(1)
```

```
                    NLAG = 24;
      ESTIMATE
                    Q = 1;
      ESTIMATE
                Q = 2;
RUN;
QUIT;


TITLE "5-Min 2nd Diff MA(1) and MA(2)";
PROC ARIMA DATA = Trts.Station_27_5Min
      PLOTS(UNPACK) = ALL;
      IDENTIFY
                    VAR  = CSF_5MinAvg(1,1)
                    NLAG = 24;
      ESTIMATE
                    Q = 1;
      ESTIMATE
                Q = 2;
RUN;
QUIT;
/*Findings*/
*1 The MA(2) model for first-difference series is the best;
*2 The MA(2) model for second-diff violates the invertibility requirement;
TITLE "5-Min 1st Diff MA(1) Outlier";
PROC ARIMA DATA = Trts.Station_27_5Min
      PLOTS(UNPACK) = ALL;
      IDENTIFY
                    VAR  = CSF_5MinAvg(1)
                    NLAG = 24;
      ESTIMATE
                Q = 1;
      OUTLIER;
RUN;
QUIT;


TITLE "5-Min 1st Diff MA(2) Outlier";
PROC ARIMA DATA = Trts.Station_27_5Min
      PLOTS(UNPACK) = ALL;
      IDENTIFY
                    VAR  = CSF_5MinAvg(1)
                    NLAG = 24;
      ESTIMATE
                Q = 2;
      OUTLIER;
RUN;
QUIT;

/*Findings*/
*1 Both the MA(1)and MA(2) model OUTLIERS: 26 AND 85;

/*------------REDO THE ANALYSIS FOR 5-MIN WITHOUT OUTLIERS---------------*/
DATA Trts.Station_27_5Min_NOL;
      SET Trts.Station_27_5Min;
      IF Traffic_Time = '30OCT2009:07:05:00.000'DT
    OR Traffic_Time = '30OCT2009:12:00:00.000'DT
```

```
    THEN DELETE;
RUN;


TITLE "5-Min 1st Diff MA(1) NOL Forecast";
PROC ARIMA DATA = Trts.Station_27_5Min_NOL
      PLOTS(UNPACK) = ALL;
      IDENTIFY
                  VAR  = CSF_5MinAvg(1)
                  NLAG = 24;
      ESTIMATE
              Q = 1;
      FORECAST
                  LEAD = 6
                  OUT = Trts.Station_27_5Min_NOL_MA1_F;

RUN;
QUIT;

TITLE "5-Min 1st Diff MA(2) NOL Forecast";
PROC ARIMA DATA = Trts.Station_27_5Min_NOL
      PLOTS(UNPACK) = ALL;
      IDENTIFY
                  VAR  = CSF_5MinAvg(1)
                  NLAG = 24;
      ESTIMATE
              Q = 2;
      FORECAST
                  LEAD = 6
                  OUT = Trts.Station_27_5Min_NOL_MA2_F;

RUN;
QUIT;
```

APPENDIX B INPUT FILES FOR QUEENSOD MODEL

Master File

```
QOD Master File - I66E OD 5:00:00
11          100
5           20
1           1
qod_050000\
qod_050000\output
qod1.txt
qod2.txt
none
none
none
f050000.txt
none
none
none
qod10.out
qod11.out
qod12.out
qod13.out
qod14.out
qod15.out
qod16.out
```

Node File

```
I66E OD Node
File
25    1.0       1.0
1     0.25      1.75    3       1       0
2     0.25      1.75    3       1       0
3     0.25      1.75    3       1       0
4     0.25      1.75    3       1       0
5     0.25      1.75    3       1       0
6     0.25      1.75    2       1       0
7     0.25      1.75    2       1       0
8     0.25      1.75    2       1       0
9     0.25      1.75    2       1       0
10    0.25      1.75    2       1       0
11    0.25      1.75    2       1       0
12    0.25      1.75    2       1       0
13    0.25      1.75    4       1       0
14    0.25      1.75    4       1       0
15    0.25      1.75    4       1       0
16    0.25      1.75    4       1       0
17    0.25      1.75    4       1       0
18    0.25      1.75    4       1       0
19    0.25      1.75    4       1       0
20    0.25      1.75    4       1       0
21    0.25      1.75    4       1       0
22    0.25      1.75    4       1       0
23    0.25      1.75    4       1       0
24    0.25      1.75    4       1       0
25    0.25      1.75    4       1       0
```

Link File

```
1.01.0 1.01.0
13 0.7788 23004 0 0 0 0 0 0 0 0 0 0 0 00000 11111
14 0.7988 23004 0 0 0 0 0 0 0 0 0 0 0 00000 11111
15 0.6 88 23004 0 0 0 0 0 0 0 0 0 0 0 00000 11111
16 0.7788 23004 0 0 0 0 0 0 0 0 0 0 0 00000 11111
17 0.8588 23004 0 0 0 0 0 0 0 0 0 0 0 00000 11111
18 1.2688 23004 0 0 0 0 0 0 0 0 0 0 0 00000 11111
19 0.5788 23005 0 0 0 0 0 0 0 0 0 0 0 00000 11111
20 0.7188 23004 0 0 0 0 0 0 0 0 0 0 0 00000 11111
21 0.5188 23004 0 0 0 0 0 0 0 0 0 0 0 00000 11111
22 0.6988 23004 0 0 0 0 0 0 0 0 0 0 0 00000 11111
23 0.6588 23004 0 0 0 0 0 0 0 0 0 0 0 00000 11111
24 0.7888 23004 0 0 0 0 0 0 0 0 0 0 0 00000 11111
25 0.4988 23004 0 0 0 0 0 0 0 0 0 0 0 00000 11111
12 0.4088 23004 0 0 0 0 0 0 0 0 0 0 0 00000 11111
6  1.0040 15001 0 0 0 0 0 0 0 0 0 0 0 00000 11111
7  1.0040 15001 0 0 0 0 0 0 0 0 0 0 0 00000 11111
16 1.0040 15001 0 0 0 0 0 0 0 0 0 0 0 00000 11111
8  1.0040 15001 0 0 0 0 0 0 0 0 0 0 0 00000 11111
9  1.0040 15001 0 0 0 0 0 0 0 0 0 0 0 00000 11111
21 1.0040 15002 0 0 0 0 0 0 0 0 0 0 0 00000 11111
10 1.0040 15001 0 0 0 0 0 0 0 0 0 0 0 00000 11111
24 1.0040 15001 0 0 0 0 0 0 0 0 0 0 0 00000 11111
11 1.0040 15001 0 0 0 0 0 0 0 0 0 0 0 00000 11111
12 1.0040 15002 0 0 0 0 0 0 0 0 0 0 0 00000 11111
```

REFERENCES

Abrahamsson, T. (1998). Estimation of Origin-Destination Matrices Using Traffic Counts - A Literature Survey. Laxenburg, Austria.

Ahmed, M. S. and A. R. Cook (1979). "Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques." Transportation Research Record **722**: 1-9.

Ahmed, S. A. and A. R. Cook (1980). "Time Series Models for Freeway Incident Detection." Transportation engineering journal of ASCE **106**(Compendex): 731-745.

Ahmed, S. A. and A. R. Cook (1980). "Time Series Models for Freeway Incident Detection." Journal of Transportation Engineering **106**: 731-745.

Alibabai, H. and H. S. Mahmassani (2008). "Dynamic origin-destination demand estimation using turning movement counts." Transportation Research Record(2085): 39-48.

Amin, M. R. and J. H. Banks (2005). "Variation in freeway lane use patterns with volume, time of day, and location." Transportation Research Record(1934): 132-139.

Antoniou, C., M. Ben-Akiva, et al. (2006). "Dynamic traffic demand prediction using conventional and emerging data sources." IEE Proceedings: Intelligent Transport Systems **153**(Compendex): 97-104.

Ashok, K. (1996). Estimation and Prediction of Time-Dependent Origin-Destination Flows. Department of Civil and Environmental Engineering. Cambridge, MA, Massachusetts Institute of Technology. **Doctor of Philosophy:** 155.

Bhattacharjee, D., K. C. Sinha, et al. (2001). "Modeling the effects of traveler information on freeway origin-destination demand prediction." Transportation Research Part C: Emerging Technologies **9**(6): 381-398.

Bikowitz, E. W. and S. P. Ross (1985). "Evaluation and Improvement of Inductive Loop Traffic Detectors." Transportation Research Record(1010): 76-80.

Bottom, J. (2000). Consistent Anticipatory Route Guidance. Department of Civil and Environmental Engineering. Cambridge, MA, Massachusetts Institute of Technology. **Doctor of Philosophy:** 251.

Bowman, A. and A. Azzalini (1997). Applied Smoothing Techniques for Data Analysis. Oxford, New York, Claredon Press.

Camus, R., G. E. Cantarella, et al. (1997). "Real-time estimation and prediction of origin-destination matrices per time slice." International Journal of Forecasting **13**(Copyright 1997, IEE): 13-19.

Carter, M., H. Rakha, et al. (1999). "Variability of traffic-flow measures across freeway lanes." Canadian Journal of Civil Engineering **26**: 270-281.

Cascetta, E. and G. Marquis (1993). "Dynamic estimators of origin-destination matrices using traffic counts." Transportation Science **27**(4): 363-373.

Chen, C., J. Kwon, et al. (2003). Detecting Errors and Imputing Missing Data for Single Loop Surveillance Systems. the 82nd Annual Meeting of Transportation Research Board. Washington D.C.

Chen, L. and A. D. May (1987). "Traffic Detector Errors and Diagnostics." Transportation Research Record **1132**: 82-93.

Chen, Y., X. Qin, et al. (2008). A Hybrid Process of Micro-Simulation and Logistic Regression for Short-term Work Zone Traffic Diversion. The 87th Annual Meeting of the Transportation Research Board. Washington D.C.

Cheu, R.-L. (1998). Freeway traffic prediction using neural networks. Proceedings of the 1998 5th International Conference on Applications of Advanced Technologies in Transportation, April 26, 1998 - April 29, 1998, Newport Beach, CA, USA, ASCE.

Cleghorn, D., F. L. Hall, et al. (1991). "Improved Data Screening Techniques for Freeway Traffic Management System." Transportation Research Record **1320**: 17-23.

Coifman, B. (1996). "New Methodology for Smoothing Freeway Loop Detector Data: Introduction to Digital Filtering." Transportation Research Record **1554**: 142-154.

Daganzo, C. (1997). Fundamentals of Transportation and Traffic Operations. Oxford, New York, Pergamon-Elsevier.

Davis, G. A. and N. L. Nihan (1991). "Nonparametric regression and short-term freeway traffic forecasting." Journal of Transportation Engineering **117**(Compendex): 178-188.

Dhingra, S. L., P. P. Mujumdar, et al. (1993). "Application of time series techniques for forecasting truck traffic attracted by the Bombay Metropolitan Region." Journal of Advanced Transportation **27**(Compendex): 227-249.

Fernandez-Moctezuma, R. J., K. A. Tufte, et al. (2007). Toward management and imputation of unavailable data in online advanced traveler information systems. 2007 IEEE Intelligent Transportation Systems Conference, 30 Sept.-3 Oct. 2007, Piscataway, NJ, USA, IEEE.

FHWA. (2005). "Focus on Congestion Relief." Retrieved August 2, 2010, from http://www.fhwa.dot.gov/congestion/.

Hellinga, B. (1994). Estimating Dynamic Origina-Destination Demands from Link and Probe Counts. Department of Civil Engineering. Kingston, Ontario, Queen's of Maryland. **Doctor of Philosophy:** 235.

Hellinga, B. R. (1994). Estimating Dynamic Origin-Destination Demands from Link and Probe Counts. Department of Civil Engineering. Ontario, Queen's University. **Doctor of Philosophy:** 211.

Hellinga, B. R. and M. Van Aerde (1998). Estimating dynamic O-D demands for a freeway corridor using loop detector data. Canadian Society for Civil Engineering - 1998 Annual Conference, June 10, 1998 - June 13, 1998, Halifax, NS, Canada, Canadian Society for Civil Engineering.

Institute of Transportation Engineers (2000). Intelligent Transportation Primer. Washington D.C, Institute of Transportation Engineers.

Ishak, S. and H. Al-Deek (2002). "Performance Evaluation of Short-Term Time-Series Traffic Prediction Model." Journal of Transportation Engineering **128**(6): 490.

Jacobson, L. N., N. L. Nihan, et al. (1990). "Detecting Erroneous Loop Detector Data in a Freeway Traffic Management System." Transportation Research Record(1287): 151-166.

Kikuchi, S. and D. Milkovic (1999). "Method to preprocess observed traffic data for consistency. Application of fuzzy optimization concept." Transportation Research Record(1679): 73-80.

Kirby, H. R., S. M. Watson, et al. (1997). "Should we use neural networks or statistical models for short-term motorway traffic forecasting?" International Journal of Forecasting **13**(Copyright 1997, IEE): 43-50.

Kwon, J., C. Chen, et al. (2004). "Statistical methods for detecting spatial configuration errors in traffic surveillance sensors." Transportation Research Record **1870**: 124-132.

Lee, S. and D. B. Fambro (1999). "Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting." Transportation Research Record **1678**: 179-188.

Levin, M. and Y. Tsao (1980). "On Forecasting Freeway Occupancies and Volumes." Transportation Research Record **773**: 47-49.

May, A. (1990). Traffic Flow Fundamentals. Englewood Cliffs, New Jersey.

Nihan, N. L. (1997). "Aid to Determining Freeway Metering Rates and Detecting Loop Errors." Journal of Transportation Engineering **123**(6): 454-458.

Nihan, N. L. and K. O. Holmesland (1980). "Use of the Box and Jenkins Time Series Technique in Traffic Forecasting." Transportation **9**: 125-143.

Nihan, N. L., Y. Wang, et al. (2006). Improving Dual-Loop Truck (and Speed) Data: Quick Detection of Malfunctioning Loops and Calculation of Required Adjustments. Seattle, University of Washington**:** 66.

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

Ortuzar, J. and L. Willumsen (1994). Modelling Transport. Chichester, New York Wiley.

Park, E. S., S. Turner, et al. (2003). "Empirical Approaches to Outlier Detection in Intelligent Transportation Systems Data." Transportation Research Record **1840**: 21-30.

Payne, J. H. and S. Thompson (2007). "Malfunction Detection and Data Repair for Induction-Loop Sensors Using I-880 Data Base." Transportation Research Record **1570**: 191-201.

Peeta, S. and I. Anastassopoulos (2002). "Automatic real-time detection and correction of erroneous detector data with fourier transforms for online traffic control architectures." Transportation Research Record **1811**: 1-11.

Ross, P. (1982). "Exponential Filtering of Traffic Data." Transportation Research Record **869**: 43-49.

Ruimin, L. and L. Huapu (2009). Combined neural network approach for short-term urban freeway traffic flow prediction. Advances in Neural Networks - ISNN 2009. 6th International Symposium on Neural Networks, ISNN 2009, 26-29 May 2009, Berlin, Germany, Springer Verlag.

Schrank, D. and T. Lomax (2009). 2009 Urban Mobility Report. College Station, Texas, Texas Transportation Institute**:** 43.

Sheffi, Y. (1985). Urban Transportation Networks: Equiblibrium Analysis with Mathematical Programming Methods. Englewood Cliffs, New Jersey, Prentice-Hall.

Smith, B. L. (1995). Forecasting Freeway Traffic Flow for Intelligent Transportation Systems Applications. Department of Civil Engineering. Charlottesville, University of Virginia. **Doctor of Philosophy**.

Szeto, W. Y., B. Ghosh, et al. (2009). "Multivariate traffic forecasting technique using cell transmission model and SARIMA model." Journal of Transportation Engineering **135**(Compendex): 658-667.

Takezawa, K. (2006). Introduction to nonparametric regression Hoboken, N.J., Wiley-Interscience.

Tao, Z., H. Lifang, et al. (2010). Nonparametric regression for the short-term traffic flow forecasting. 2010 International Conference on Mechanic Automation and Control Engineering (MACE), 26-28 June 2010, Piscataway, NJ, USA, IEEE.

Taylor, C. and D. Meldrum (1995). Freeway traffic data prediction using neural networks. Pacific Rim TransTech Conference. 1995 Vehicle Navigation and Information Systems Conference Proceedings. 6th International VNIS. A Ride into the Future, 30 July-2 Aug. 1995, New York, NY, USA, IEEE.

Turner, S. (2004). "Defining and measuring traffic data quality: White paper on recommended approaches." Transportation Research Record **1870**: 62-69.

Van Aerde, M., H. Rakha, et al. (2003). "Estimation of Origin-Destination Matrices: Relationship between Practical and Theoretical Considerations." Transportation Research Record **1831**: 122-130.

Van Der Voort, M., M. Dougherty, et al. (1996). "Combining Kohonen maps with ARIMA time series models to forecast traffic flow." Transportation Research Part C (Emerging Technologies) **4C**(Copyright 1997, IEE): 307-318.

Vanajakshi, L. and L. R. Rilett (2004). "Loop detector data diagnostics based on conservation-of-vehicles principle." Transportation Research Record(1870): 162-169.

Weijermars, W. A. M. and E. C. Van Berkum (2006). "Detection of invalid loop detector data in urban areas." Transportation Research Record **1945**: 82-88.

Whittaker, J., S. Garside, et al. (1997). "Tracking and predicting a network traffic process." International Journal of Forecasting **13**(Copyright 1997, IEE): 51-61.

Yang, H., T. Sasaki, et al. (1992). "Estimation of origin-destination matrices from link traffic counts on congested networks." Transportation Research, Part B: Methodological **26**(6): 417-417.

Zhou, X. (2004). Dynamic Origin-Destination Demand Estimation and Prediction for Off-line and On-line Dynamic Traffic Assignment Operation. Department of Civil and Environmental Engineering. College Park, Maryland, University of Maryland. **Doctor of Philosophy:** 174.