12-2017

Real-time Traffic Flow Detection and Prediction Algorithm: Data-Driven Analyses on Spatio-Temporal Traffic Dynamics

Bumjoon Bae
University of Tennessee, bbae1@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Bumjoon Bae entitled "Real-time Traffic Flow Detection and Prediction Algorithm: Data-Driven Analyses on Spatio-Temporal Traffic Dynamics." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Civil Engineering.

Lee D. Han, Major Professor

We have read this dissertation and recommend its acceptance:

Hamparsum Bozdogan, Christopher Cherry, Hyun Kim

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Real-time Traffic Flow Detection and Prediction Algorithm: Data-Driven Analyses on Spatio-Temporal Traffic Dynamics

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Bumjoon Bae

December 2017

DEDICATION

To my wife and son, Lina Ryu and Jiho B. Bae, for all their endless love and support.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my advisor, Dr. Lee D. Han for his support and guidance during my graduate studies. I was very fortunate to have him as my advisor who continually and convincingly conveyed keen insight with regard to my research. Without his encouragement, I would never finish my dissertation successfully. I would also like to thank my statistics advisor, Dr. Hamparsum Bozdogan, and my committee members, Dr. Christopher Cherry and Dr. Hyun Kim for serving on my committee, as well as Dr. Asad Khattak for helping me improve my study.

I am truly grateful to Hyeonsup Lim for his endless support and excitement about research and life in Knoxville as my mentor. I will definitely miss the days in Knoxville with him.

Last, but not least, I would like to thank all my lab mates: Stephanie Hargrove, Jianjiang Yang, Yang Zhang, Yuandong Liu, Zhihua Zhang, Pankaj Dahal, Brandon Whetsel, Brandon Worley, Kwaku Boakye, Meng Zhang, Ziwen Ling, Jun Liu, Ranjit Khatri, Nirbesh Dhakal, Ali Boggs, Behram Wali, Xiaobing Li, Mohsen Kamrani, Nima Hoseinzadeh and many others.

ABSTRACT

Traffic flows over time and space. This spatio-temporal dependency of traffic flow should be considered and used to enhance the performance of real-time traffic detection and prediction capabilities. This characteristic has been widely studied and various applications have been developed and enhanced. During the last decade, great attention has been paid to the increases in the number of traffic data sources, the amount of data, and the data-driven analysis methods. There is still room to improve the traffic detection and prediction capabilities through studies on the emerging resources. To this end, this dissertation presents a series of studies on real-time traffic operation for highway facilities focusing on detection and prediction.

First, a spatio-temporal traffic data imputation approach was studied to exploit multi-source data. Different types of kriging methods were evaluated to utilize the spatio-temporal characteristic of traffic data with respect to two factors, including missing patterns and use of secondary data. Second, a short-term traffic speed prediction algorithm was proposed that provides accurate prediction results and is scalable for a large road network analysis in real time. The proposed algorithm consists of a data dimension reduction module and a nonparametric multivariate time-series analysis module. Third, a real-time traffic queue detection algorithm was developed based on traffic fundamentals combined with a statistical pattern recognition procedure. This algorithm was designed to detect dynamic queueing conditions in a spatio-temporal domain rather than detect a queue and congestion directly from traffic flow variables. The algorithm was evaluated by using various real congested traffic flow data. Lastly, gray areas in a decision-making process based on quantifiable measures were addressed to cope with uncertainties in modeling outputs. For intersection control type selection, the gray areas were identified and visualized.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Traffic congestion on a highway is one of the most interesting phenomena in traffic management and operations. A great deal of effort has been made to develop and enhance solutions to cope with traffic congestion.

Assessing current traffic conditions and accurately predicting the future in real time are essentials that have not been definitively resolved. Advancement in these capabilities is vital for fast decision making, timely responses, and appropriate proactive traffic controls to mitigate the impact of congestion on traffic flow.

One purpose of traffic flow analysis is to understand the continuous movement of traffic over time and space. The traffic variables including speed, density, and traffic volume depend on a spatio-temporal domain. Without considering this feature of the data in an analysis, one can make only limited inferences. Thus, more efforts to exploit the spatio-temporal dependency in traffic flow studies are desirable.

Traffic data collected from intelligent transportation systems (ITS) and mobile devices have increased rapidly with significant progress in computing capabilities and data-driven analysis methods. Data sources include conventional traffic data from detectors as well as location-based data from cellphones, car navigation devices, and multiple sensors embedded in connected and autonomous vehicles. In order to conduct further analysis using these data, one must address a missing data issue. Missing data appears frequently in a real traffic data collection process. This is mainly because collecting data from transportation systems is different from collecting it under well-controlled experimental conditions. If a great deal of data is missing, it can lead to an erroneous analysis.

For traffic flow analysis, the wide variety of sources provides data that represent traffic flow conditions. Speed is one essential type of these data.

Speed is a fundamental variable of traffic flow, and is frequently used for highway capacity analysis, although it is not directly used as a level of service measure. Various traffic phenomena occur in a highway system due to weaving, merging, diverging, and other traffic events that are often measured and explained by speed. Therefore, traffic flow analysis using speed data can provide important information for detecting and predicting traffic conditions.

This dissertation presents studies focused on the development of a traffic flow detection and prediction framework that uses data-driven approaches to support the proactive traffic controls and operations for highways. The dissertation compiles four research papers in the following chapters. These chapters are organized in a journal article format because each chapter is either published, submitted, or to be submitted.

- Chapter I proposes and evaluates spatio-temporal cokriging methods for missing data imputation in spatio-temporal domain. Different missing data patterns and use of secondary data are considered for enhancing imputation accuracy.
- Chapter II proposes a short-term traffic speed prediction algorithm. A nonparametric time series analysis method with a data dimension reduction technique is evaluated for short-term prediction and compared with a parametric model.
- Chapter III presents a real-time traffic queue detection algorithm based on traffic flow fundamentals using traffic detector data. The queue detection and additional shock wave analysis results are provided using real traffic data.
- Chapter IV introduces gray areas in a decision-making process using a quantifiable performance measure to address uncertainties in modeling output. The gray areas are identified and visualized in a case study of intersection control type selection.

CHAPTER I

MISSING DATA IMPUTATION FOR TRAFFIC SPEED USING SPATIO-TEMPORAL COKRIGING

This chapter presents a modified version of a research paper by Bumjoon Bae, Hyun Kim, Hyeonsup Lim, Yuandong Liu, Lee D. Han, and Phillip B. Freeze.

Abstract

Modern transportation systems rely increasingly on the availability and accuracy of traffic detector data to monitor traffic operational conditions and assess system performance. Missing data, which occurs almost inevitably for a number of reasons, can lead to suboptimal operations and ineffective decisions if not remedied in a timely and systematic fashion through data imputation. A review of the literature suggests that most traffic data imputation studies considered the temporal continuity of the data but often overlooked the spatial correlations that exist. Few of the studies explored the randomness of the patterns of the missing data. Therefore, this paper proposes two cokriging methods that exploit the existence of spatiotemporal dependency in traffic data and employ multiple data sources, each with independently missing data, to impute high-resolution traffic speed data under different data missing pattern scenarios. The two proposed cokriging methods, both using multiple independent data sources, were benchmarked against a classic ordinary kriging method, which uses only the primary data source. An array of testing scenarios were designed to test these methods under different missing rates (10~40% data loss) and different missing patterns (random in time and location, random only in location, and non-random blocks of missing data). The results suggest that using multiple data sources with the spatiotemporal simple cokriging method effectively improves the imputation accuracy if the missing data were clustered, or in blocks. On the other hand, if the missing data were randomly scattered in time and location, the classic ordinary kriging method using only the primary data source can be more effective. Our study, which employs empirical traffic speed data from radar

detectors and vehicle probes, demonstrates that the overall predictions of the kriging-based imputation approach are accurate and reliable for all combinations of missing patterns and missing rates investigated.

Introduction

Traffic detector data collected from transportation facilities are essential inputs for modern transportation systems to monitor traffic conditions and assess system performance. A challenge for using the data is 'missingness' in the data collection processes of the systems [1, 2]. This includes (but is not limited to) the malfunctions of hardware or software, communication network problems, restricted power supply conditions, scheduled maintenance, and so on. As Orchard and Woodbury [3] remarked, it is obvious that not to have missing data is the best way to address the missing data issue; however, this ideal circumstance rarely happens.

The effects of missing data and imputation methods have been examined in other disciplines, such as statistics, sociology, and epidemiology because analysis results are considered rough when data are missing [1]. Unfortunately, this issue has not been well addressed in transportation studies [4, 5]. Measuring the effects of missing data and treatments to impute them are rarely investigated, even though the issue of handling missing data has been addressed to some degree in transportation modeling. Meanwhile, the need to measure the performance of transportation systems such as delay, travel time reliability, and emissions has been underlined in transportation systems management and operations. In this context, the appropriate methods to impute missing data should be explored, otherwise, the results of such performance measures will be biased.

The effects of missing traffic flow data on transportation modeling and prediction can be divided into two categories [6]: First, it causes information loss for certain locations and time periods, which may be important to the objective of an analysis in transportation modeling and prediction. For instance, if traffic speed and traffic volume data are missing for a severely congested road segment during peak hours, the total vehicle emission will be underestimated. Second, it causes statistical information loss. In general, a sample size that is smaller due to missing data, i.e., smaller degrees of freedom, may lead to overfitting problems in the modeling process. More importantly, underlying assumptions of statistical methods used in an imputation analysis are violated by different missing patterns, resulting in biased solutions.

Therefore, to avoid erroneous statistical inference, understanding missing patterns and missing mechanisms from the datasets used in a statistical analysis is as important as determining how to sample from a population. Rubin [7] points out that distributional inferences on the parameters of data are generally conditional on the observed missing patterns. According to recent works by Buuren [1] and Carpenter and Kenward [2], a typology of missing patterns associated with the impact on statistical analysis are identified with three types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Few of previous studies explored the randomness of the patterns of the missing traffic flow data, but not fully investigated these missing patterns [8-11].

Imputation of traffic data such as volume, speed, and occupancy collected from traffic detectors aims to estimate the unobserved value at a specific location and time to improve the accuracy of further analyses (traffic speed prediction, traffic incident detection, and so on). Recent transportation studies paid attention to a geostatistical approach, called kriging, to estimate or predict traffic variables for unobserved locations [12-16]. Considering that traffic data have spatiotemporal dependency, kriging has an advantage over other statistical

approaches for improving imputation accuracy. This is because the method takes the observed neighboring data correlated with a missing value into account in space-time dimension. A recent kriging study extends the modeling dimension from a single spatial dimension to a spatiotemporal dimension to impute traffic speed data, arguably suggesting that spatiotemporal-kriging (ST-Kriging) outperforms the historical average and k-nearest neighborhood (KNN) methods [17].

The goal of this study is to extend the ST-Kriging approach to a multivariate framework (called spatiotemporal cokriging) for imputing high-resolution traffic speed data collected from Remote Traffic Microwave Sensors (RTMS) on highways. As cokriging is inherently the multivariate extension of kriging [18], cokriging needs to input the secondary variables to complement the observed neighboring values of the primary variable to predict the value of a primary variable at a new location. The secondary variables are spatially correlated with the primary variable. Because available traffic data resources are abundant, using the information from multiple data sources is anticipated to improve the imputation results of the spatio-temporal cokriging approach. The effectiveness of cokriging relies on the pattern of missing data. To address this issue, we investigated the prediction performance of three different kriging methods based on three missing patterns (MCAR, MAR, and MNAR) in the traffic speed data.

The next section presents a comprehensive literature review on imputation techniques and kriging in transportation studies and describes the data used in this study. The following section explains the kriging and cokriging methods. The last two sections provide a case study result of applying the spatiotemporal cokriging approach to impute traffic flow speed data, then conclusions follow.

Literature Review

Missing data imputation methods can be either single imputation or multiple imputation [1, 2]. Hot-deck, average, and regression are commonly used as single imputation methods. Most of the imputation studies in transportation examine single imputation methods because of their fast-computational speed for real-time analysis. The historical average, expectation maximization (EM) algorithm [4], pairwise regression [19], moving average, ARIMA, and regression model with genetic algorithm [20] have been explored for imputing five or 10 minutes loop detector data.

In contrast, multiple imputation methods overcome the drawback of single imputation methods that derive standard errors of parameter estimates that are too small. This type of imputation generates multiple imputed datasets and estimates model parameters, then pools the estimates as a single value. Thus, it can deal with the inherent uncertainty of the imputations [1]. Ni and Leonard Ii [5] proposed a multiple imputation approach employing a Bayesian network and Markov chain Monte Carlo (MCMC) technique with 20-second detector data. The imputation method can account for the correlations between and within variables by using a Bayesian network to produce unbiased estimates and confidence intervals of the results from the MCMC. However, these studies exploited only time series information of the traffic data at the location of interest or the traffic data of closest surrounding detectors, which are selected arbitrarily on the basis of spatiotemporal relationship assumed in advance.

Recent studies have focused more on both the spatial context of traffic data as well as temporal patterns [9, 11, 21]. Clearly, analyzing traffic dynamics in the context of space and time is useful since traffic status evolves in the spatiotemporal domain. Thus, efforts have been made to visualize and analyze traffic data projecting in three or more dimensions, including space and time [21, 22]. Spatiotemporal properties of traffic detector data were also explored using a

cross-correlation analysis [23]. The results can be used to identify the influential area of a missing data point in a spatiotemporal domain.

Kriging is a well-known geostatistics interpolation method developed by D. G. Krige [24] to estimate a value at an unobserved location from observations at nearby locations. Previous transportation studies exploring kriging methods were focused on estimating annual average daily traffic (AADT) for unobserved locations. Eom, Park [12] employed kriging to impute missing AADT data. In comparison with an ordinary least square (OLS) regression model, the kriging approach predicted AADT more accurately. Kriging was also used to predict future AADT with temporal extrapolation by OLS regression [13]. The study showed that kriging approaches perform better for road sections with moderate-to-high traffic volumes. Under low traffic demand conditions, the proposed kriging method overestimates AADT. Since the spatial covariance considered in kriging is based on Euclidean distance between two locations, Zou, Yue [14] proposed an approximated road network distance, which is a Euclidean distance and approximately equal to the road network distance using isometric embedding theory. Therefore, the metric can be used for the traditional kriging and its covariance function. In comparison with the Euclidean distance, the proposed distance metric performed better for interpolating travel speed using local universal kriging, especially for a region with a complex road network structure. However, Selby and Kockelman [15] showed that using network distances, instead of Euclidean distance, did not significantly improve the prediction performance of a universal kriging method for AADT prediction. Shamo, Asa [16] compared three different kriging methods—simple kriging (SK), ordinary kriging (OK), and universal kriging (UK)—to predict AADT in Washington State in U.S. over a period of three years. The result showed that there is no superior kriging method from year to year due to the dynamic nature of traffic volume. Meanwhile, there are attempts to extend the spatial analysis dimensions of kriging to a spatiotemporal domain to better capture the spatiotemporal properties of data in

various disciplines [17, 25, 26]. However, there is virtually no available literature associated with imputation and prediction modeling. Given the theoretical advantage of cokriging, applying cokriging method to impute missing traffic data is promising in case any secondary data that are highly correlated with a primary data are available. Since multiple traffic speed data sources are available, it is worthwhile to explore the applicability and performance of cokriging for traffic data imputation.

Data Description

The primary data to be imputed in this study was obtained from the RTMS, monitored by roadside sensors in the Knoxville urban area; Knoxville is the third largest city in Tennessee. There are more than 200 detector stations for both directions on the interstates, including two major highways in the Knoxville region, I-40 and I-75. Figure 1-1 shows the selected RTMS station locations together with the station ID labels. Twenty-eight stations in the 13.6 mile-long eastbound I-40 segment, ranging from mile marker 374.2 (west end) to 387.8 as (east end), were selected because it is a major city corridor. Along with Figure 1-1, Table 1-1 summarizes the selected RTMS stations that are aligned to 12 links of the secondary dataset, called HERE on I-40 Highway in Knoxville, TN. There were 5,881 cases where both speeds were collected at the same spatiotemporal point.

RTMS collects traffic count, speed, and occupancy information for each lane every 30 seconds. Speed is the essential variable for measuring the performance of the highway system in terms of travel time reliability, delay, emissions, and so on. To explore the cokriging approach for data imputation, the five-minute average speed data for 24 hours were collected on December 1, 2015. This gave a total of 288 observations for each station if no missingness

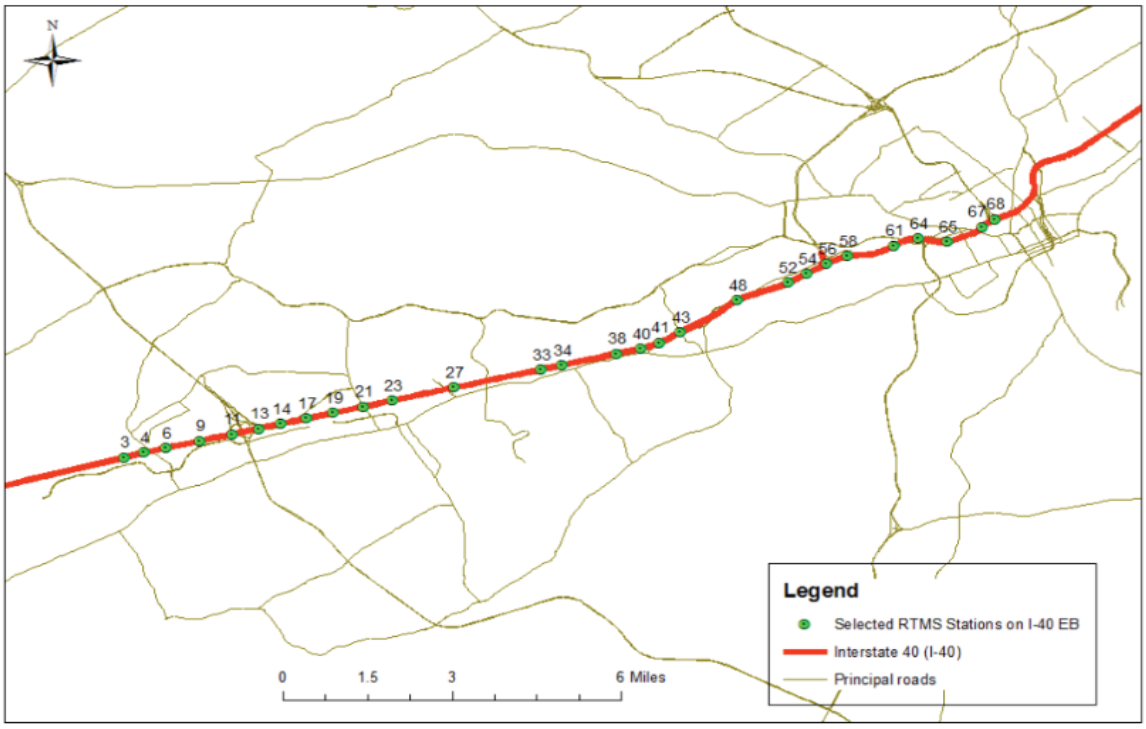


Figure 1-1 Map of the selected RTMS stations on I-40 eastbound in Knoxville.

Table 1-1 Description of RTMS stations and corresponding HERE links.

RTMS			HERE	
Station ID	Mile Marker	Direction	Link ID	Link Length (mile)
3	374.2	Eastbound	121P04124	2.0
4	374.5	Eastbound		
6	374.9	Eastbound		
9	375.4	Eastbound	121P04125	1.2
11	375.9	Eastbound		
13	376.2	Eastbound		
14	376.6	Eastbound	121P04126	0.5
17	377.1	Eastbound	121P04127	1.3
19	377.5	Eastbound		
21	378.0	Eastbound		
23	378.4	Eastbound	121P04128	0.5
27	379.2	Eastbound	121P04130	0.8
33	380.4	Eastbound	121P04131	1.2
34	380.7	Eastbound		
38	381.5	Eastbound	121P04132	2.3
40	381.9	Eastbound		
41	382.2	Eastbound		
43	382.6	Eastbound		
48	383.6	Eastbound	121P04133	1.3
52	384.4	Eastbound		
54	384.7	Eastbound	121P04144	0.8
56	385.1	Eastbound		
58	385.4	Eastbound		
61	386.1	Eastbound	121P04146	0.5
64	386.5	Eastbound		
65	387.0	Eastbound		
67	387.5	Eastbound	121P04149	0.2
68	387.8	Eastbound		

occurred. The main reason to use five-minute data is for consistency with the aggregation scale of the secondary data used in this study. Note that the raw data were collected in 30 seconds interval; however, it was aggregated to remove the effect of unnecessary noise as it is prone to a biased distribution for imputation in our study. In our examination, the aggregation of the raw data at five-minute intervals was considered a reasonable time span for imputation. The secondary data used in the cokriging method called HERE is a commercial link-based speed dataset collected mainly from probe vehicles. As mentioned, this data contains traffic speeds averaged in five minutes, totaling 288 observations per day for each road link. The HERE data were obtained at the same road segments on the same date of the RTMS dataset. Note that the RTMS dataset is point-based while the HERE dataset is link-based. Thus, to match the RTMS station locations with the links of HERE, a geographic information system (GIS) tool was used for the data matching allocation.

In each dataset, the number of complete data points for a day should be 8,064 (=28 stations $\times$288 per day for five-minute interval). However, the obtained RTMS dataset had 6,954 observations and the HERE dataset contained 6,817 observations, presenting the original missing rates of both RTMS and HERE samples in this study are 13.8% and 15.5%, respectively. Figure 1-2 shows the scatter plots of the RTMS (Figure 1-2(a)) and HERE (Figure 1-2(b)) observations in spatiotemporal dimension. Most of the missing values in the HERE data are observed at nighttime because HERE data are collected from probe vehicles that operate mostly in the daytime. Nevertheless, the HERE data are still useful for daytime analyses. In terms of missingness in the RTMS data, the missing pattern is a random pattern.

Given the obtained dataset, two assumptions are made to use the RTMS and HERE data together. First, five-minute aggregated data from the original 30-second RTMS are used for the consistency of the temporal scale of HERE. Consequently, the changes within five minutes of the raw RTMS data are

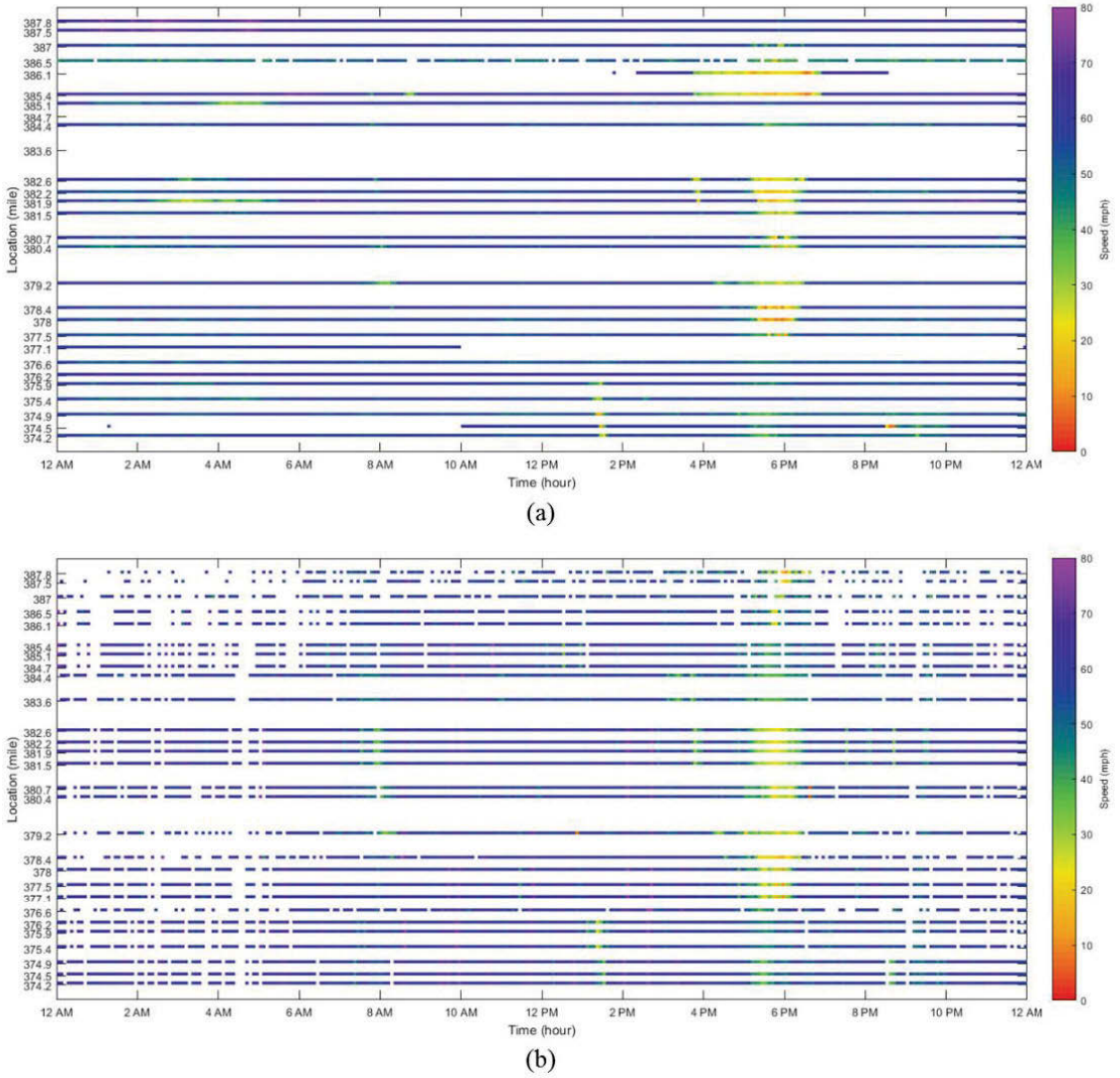


Figure 1-2 Collected five-minute average speed points: (a) RTMS and (b) HERE. Each dot represents an observation point in spatiotemporal dimension.

smoothed. This pre-processing has an advantage because the aggregation can reduce the impact of the noise in the raw data and improve computation efficiency [5]. Second, the resolution and accuracy of the secondary data, HERE are assumed to be sufficient to explain a part of the variation in the primary data within the proposed cokriging imputation framework. It is expected that the HERE data be helpful to impute missing RTMS values because the completeness rate of HERE is well established during daytime collection.

Methodology

Since kriging was first developed as an interpolation technique for geographical surfaces, it has become a representative geostatistical approach to predict an unknown value at an unobserved location by adapting various statistical assumptions and conditions in the modeling and has further advanced to different kriging methods. The interpolation was formulated as a weighted sum of the values of their known neighbors. In this study, the speed at an unobserved location is estimated using three different kriging methods: ordinary kriging (OK), ordinary cokriging (OCK), and simple cokriging (SCK).

OK is the most commonly used kriging method [27]. With local second-order stationarity assumption, OK is known as the best linear unbiased estimator (BLUE). In this study, the speed at an unobserved location $\boldsymbol{s}_0$ is calculated from a linear combination of the observed speed $V(\boldsymbol{s}_\alpha)$ at neighboring locations and its weight λ_α:

$$V^*(\boldsymbol{s}_0) = \sum_{\alpha=1}^{h} \lambda_\alpha V(\boldsymbol{s}_\alpha) \quad (1)$$

where, $\boldsymbol{s}_\alpha$ is a vector of the location where an observed RTMS speed V is placed on a spatiotemporal plane, $\boldsymbol{s}_0$ is a vector of the location where unobserved RTMS

speed V^* will be predicted, and n is the number of observed locations used for prediction. To obtain an unbiased estimate, the following constraint is added:

$$\sum_{\alpha=1}^{h} \lambda_\alpha = 1 \tag{2}$$

OCK is a multivariate extension of OK. Using OCK, secondary variables can be added to predict a primary variable. OCK can be expressed as follows:

$$V_{i_0}^*(\boldsymbol{s}_0) = \sum_{i=1}^{K} \sum_{\alpha=1}^{h_i} \lambda_\alpha^i V_i(\boldsymbol{s}_\alpha) \tag{3}$$

where, $V_{i_0}^*$ is unobserved speed of a primary variable, RTMS to be predicted and K is the number of variables, and h_i is the number of observed locations of ith variable. A similar but conditional constraint of OK is added in OCK:

$$\sum_{\alpha=1}^{h_i} \lambda_\alpha^i = \begin{cases} 1 & \text{if } i = i_0 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Then, the main difference between OCK and SCK is how the mean value is specified for interpolation. A constant *global* mean is used in SCK, while the *local* mean is used in OCK, which varies depending on each set of neighboring data points. Therefore, the accuracy of the OCK-based prediction could decrease when no neighborhood data are available. Under such conditions, SCK is more useful since an estimation of a primary variable can be calibrated without having neighboring primary data [27]. SCK is expressed as:

$$V_{i_0}^*(\boldsymbol{s}_0) = m_{i_0} + \sum_{i=1}^{K} \sum_{\alpha=1}^{h_i} \lambda_\alpha^i [V_i(\boldsymbol{s}_\alpha) - m_i] \tag{5}$$

where, m_{i_0} is the global mean of a primary variable, and m_i is the global mean of ith variable.

In kriging, the spatial dependency between two locations is analyzed by covariance or semivariogram where Euclidean distance between the observed data at any pair of locations is used to generate a best-fit semivariogram. Complete details about the semivariogram functions and underlying assumptions are available in Cressie [28], Eom, Park [12], and Zou, Yue [14]. As shown in Figure 1-2, time and space are represented with two dimensions. Notice that a road section is represented as a straight line in the Y-axis, and the observed locations are placed on the line scaled by their mile marker. Each data point was mapped on the grid in which each cell size is five-minute by 0.1-mile. Similar to what Zou, Yue [14] did, this design makes the network distance equal to the Euclidean distance, allowing the computation of spatial dependency to be more accurate and the visualization of results more effective.

Figure 1-3 presents the procedure of the analysis design: (a) collecting and preprocessing RTMS and HERE data, including map matching and data extraction and aggregation; (b) analyzing the correlation between RTMS and HERE speed data, which is for verifying that HERE is an appropriate secondary variable; (c) generating three missing patterns in the collected RTMS data; (d) creating experimental semivariogram of both data and fitting theoretical semivariogram; (e) predicting the missing RTMS speeds and mapping the spatiotemporal distribution; and (f) evaluating the accuracy of the results of OK, OCK, and SCK given the missing patterns.

Analysis Results

Correlation analysis

In order to justify the use of HERE as the secondary variable for imputing the RTMS data using cokriging, we carried out correlation analysis between two

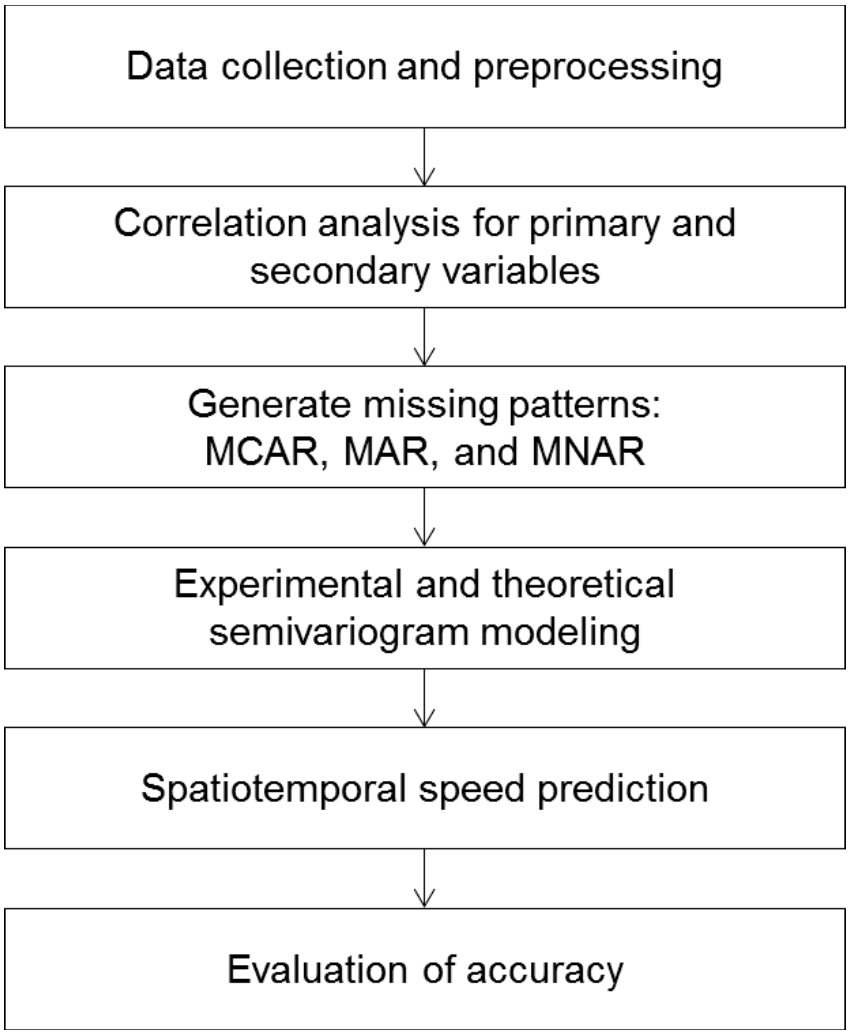


Figure 1-3 Research design.

datasets. Pearson correlation coefficient (*r*) was used as Eq. (6).

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \quad (6)$$

where, x is RTMS, y is HERE, n is sample size (*n* = 5,881), and s is a sample standard deviation of each data. As shown in Figure 1-4(a), the correlation between RTMS and HERE is 0.46. However, by taking log-transformation in Figure 1-4(b), the correlation is improved to *r* = 0.51. Empirically, this correlation is enough to justify taking the additional complexity of cokriging into account since it is known that cokriging results in better predictions than ordinary kriging if the correlation between two variables exceeds 0.5 and when a secondary variable is over-sampled [29, 30]. Most of the speed observations in both data are near 60 miles per hour (mph) and they seem to have a relatively low correlation, i.e., the cluster near 60 mph has a circular shape in Figure 1-4(a). This is mainly because of the resolution difference between both data. In other words, one HERE link covers up to 4 RTMS stations in this study. However, low speed below the free flow speed of near 60 mph is generally more important in an analysis for traffic flow since congestion represented by low speed is one of the most interesting phenomena in transportation studies. In that perspective, the low speed observations in both data show a linear relationship, which supports that HERE is an appropriate secondary data for imputing the missing RTMS data.

Design of scenarios for missing patterns

The definitions of three missing patterns – MCAR, MAR, and MNAR, are as follows: (a) the probability of a value missing at a certain location and time is completely independent in MCAR. (b) In MAR, missingness is dependent on a certain condition, but independent within the condition [1]. (c) MNAR represents the pattern that a missing mechanism is neither MCAR nor MAR. Using three

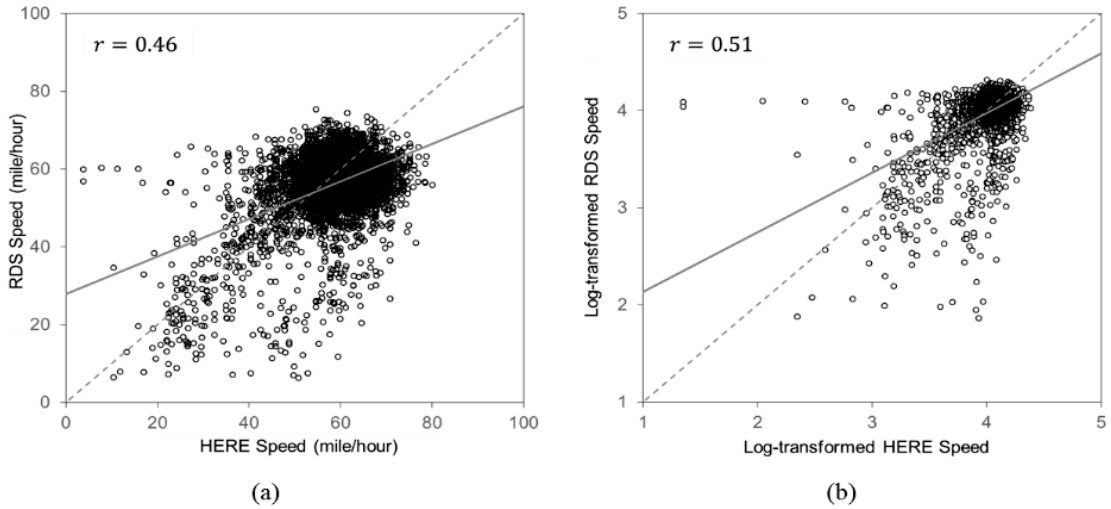


Figure 1-4 Scatter plots between RTMS and HERE: (a) original speed, and (b) log-transformed speed.

missing patterns, a total of 12 missing scenarios were generated by four missing rates from 10% to 40% in increments of 10%. In MCAR scenarios, a portion of the individual data points was removed completely at random in time and location. In MAR scenarios, series of the data points were removed for randomly selected stations and time periods to satisfy the condition that the missingness is dependent on time, but not on locations. In MNAR scenarios, a set of data points in a block was eliminated from original data to generate a different pattern compared to MAR and MCAR. Figure 1-5 shows three examples of the RTMS scatter plot given missing patterns with a scenario of 10% missing rate.

Semivariogram modeling

Using the Geostatistical Analyst tool in ArcGIS 10.4, experimental semivariograms were computed, then theoretical semivariograms were estimated for OK, OCK and SCK. Note that no dominantly superior semivariogram model has been suggested for each of the three kriging methods for traffic flow data prediction and imputation in existing literature [16, 31], implying the best fitted semivariogram model needs to be designed depending on the data used in the cases. However, recent works by Shamo, Asa [16] and Yang [31] argued that spherical and exponential models could outperform others with traffic flow data from their empirical observations. In our study, the spherical model was fitted best in our case, thus, it was applied for three kriging methods to maintain consistency in comparing them.

The HERE speed was used as the secondary variable in both cokriging methods (OCK and SCK). A theoretical semivariogram model measuring spatial dissimilarity of any pair of observations consists of three parameters: nugget (the minimum estimate of error), sill (the maximum dissimilarity), and range (the distance to reach to sill). Note that it is necessary to find the best fitting semivariogram for both data to set the equation of selected kriging methods that are best solvable in imputation. Figure 1-6 shows the best-fitted theoretical

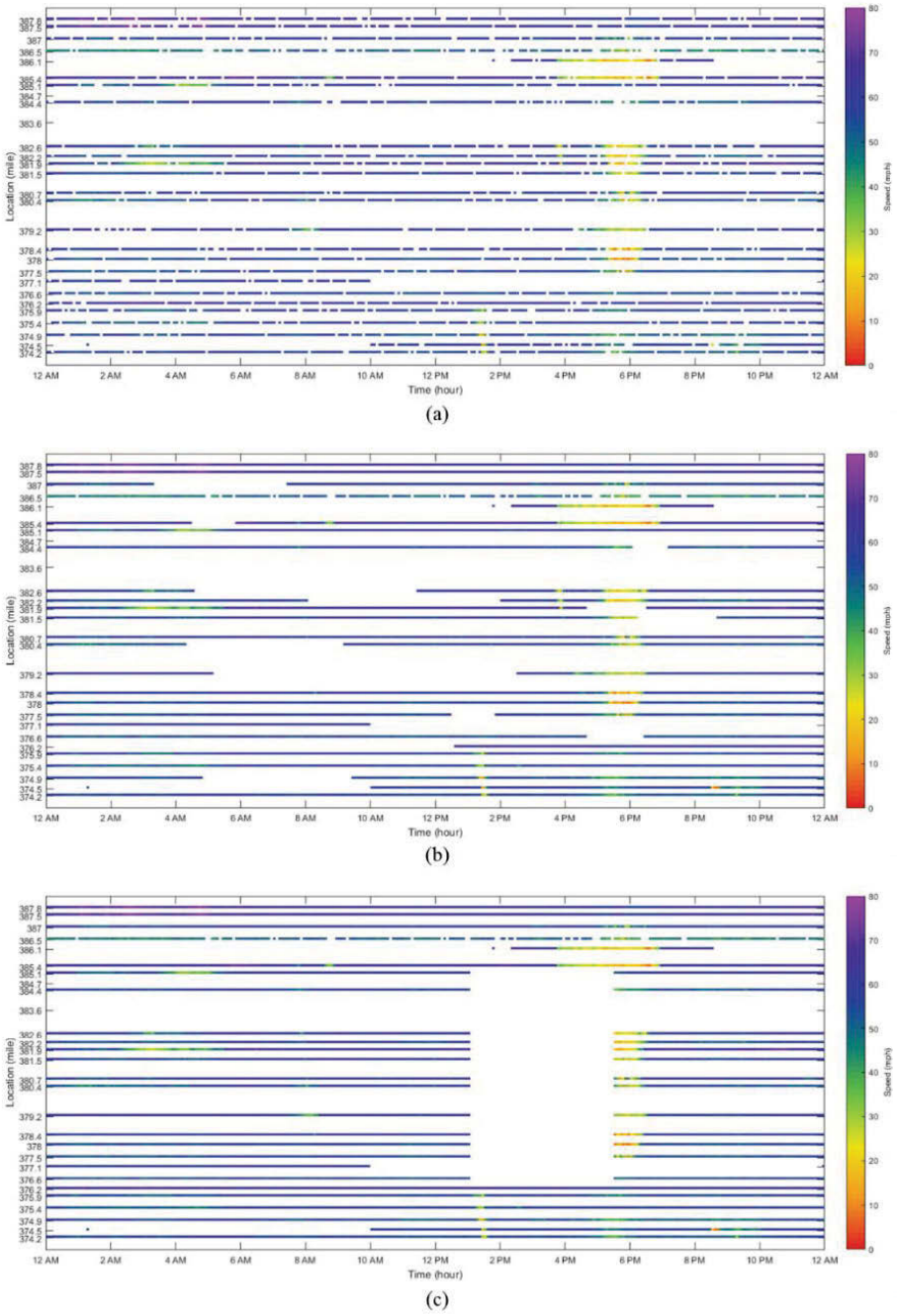


Figure 1-5 Missing scenario plots with 10% missing rate: (a) MCAR, (b) MAR, and (c) MNAR. Each dot represents observation point in spatiotemporal dimension.

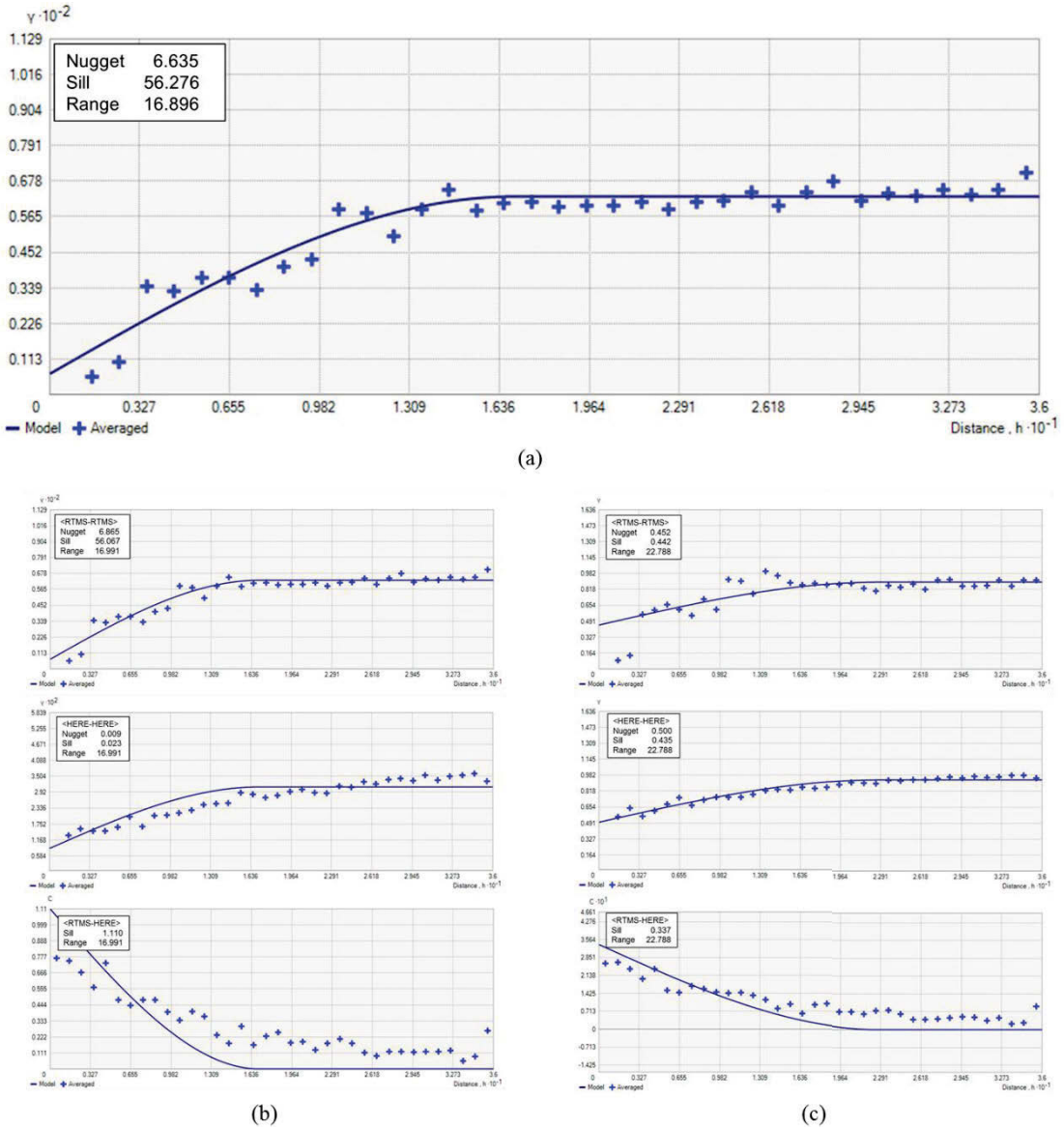


Figure 1-6 Theoretical semivariograms: (a) OK, (b) OCK, and (c) SCK.

semivariograms of the three kriging methods. The process to identify the best-fitting theoretical semivariogram over experimental semivariogram is difficult if greater variability is present in the pattern of binned cloud [27, 28]. To tackle this issue, the log-transformation was applied to HERE data in OCK and the RTMS and HERE data were transformed as normal scores in SCK. This is because simple kriging requires the assumption that the true mean of data must be known, which is identified as the best theoretical semivariogram for SCK [32].

Evaluation of the results

Given the missing patterns scenarios, a total 36 RTMS speed surfaces were generated to evaluate the imputation performance of the three kriging methods. Figure 1-7 is provided as reference patterns of the RTMS and HERE datasets (Figure A-1, Figure A-2, and Figure A-3 show the imputation results in Appendix). The removed observations in each analysis were used as ground truth for individual missing values. Note that the original missing RTMS values were not accounted for in the evaluation because their true values are not available. In order to evaluate imputation performance of OK, OCK, and SCK, *mean absolute error* (MAE) and *mean absolute percentage error* (MAPE) were used. Both measurements are formulated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|V_i - \hat{V}_i\right| \tag{7}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|V_i - \hat{V}_i\right|}{V_i} \times 100 \tag{8}$$

where, V_i is the ith observed RTMS speed and $\hat{V}_i$ is the ith predicted RTMS speed.

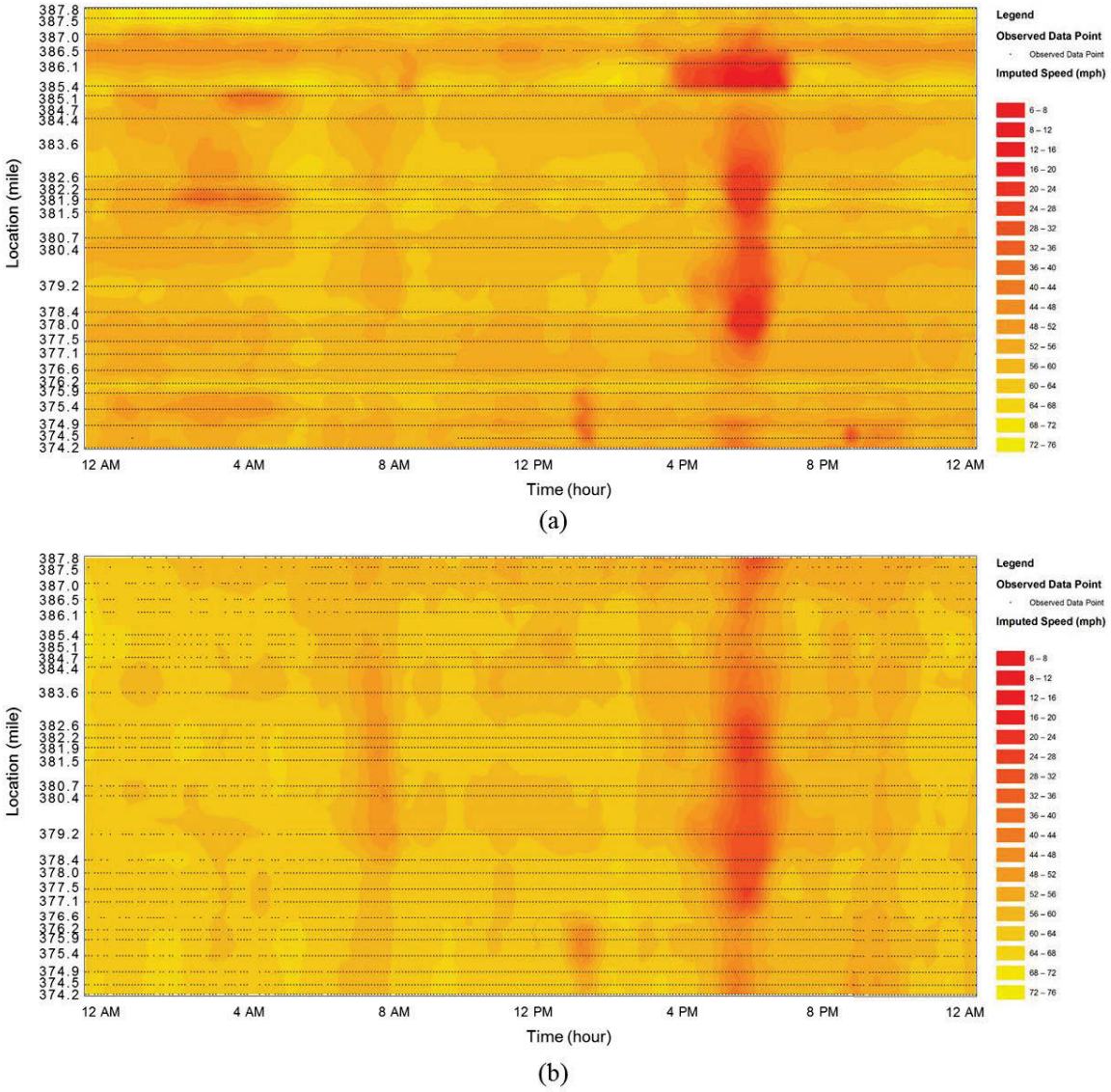


Figure 1-7 Imputed Speed using OK without missing values: (a) RTMS and (b) HERE.

The six plots in Figure 1-8 present the prediction performance of the three kriging methods under different missing data patterns and rates (The same results are presented in Table A-1 in Appendix).

For the MCAR scenarios depicted in Figure 1-8(a) and Figure 1-8(b), where data are missing completely at random, ordinary kriging (OK) clearly outperforms the others. Since plentiful of neighboring RTMS data are present near each missing point in both space and time domains in the MCAR scenario, the availability of neighboring HERE data points does not contribute meaningfully to the imputation effort. Despite the different performances among the three kriging methods, the error range of 1.9 – 3.2 mph, which can be ignored for imputation, confirms that kriging is an effective tool for missing data imputation if a missing pattern presents with a form of MCAR. Figure 1-8(a) and Figure 1-8(b) indicate that kriging-based imputation can provide reliable results for the MCAR pattern. The performance of three kriging approaches is consistently stable with varying missing rate. Notice that a single data based kriging imputation (OK) provides a "good" result when supplementary data would play as a role of noise in MCAR. However, it is worth noting that the MCAR pattern is less likely to occur in real traffic detector data.

The prediction errors of the MAR scenarios are shown in Figure 1-8(c) and Figure 1-8(d); the mean error (left) and mean percentage error (right) on average are 5.4 mph and 11.2%, respectively, both of which are higher than the ranges of MCAR results. The main reason for this result is that the missing values are more clustered as a form of time series in the scenarios compared to MCAR. This result concurs with the argument that temporal dependency of traffic detector

data is stronger than spatial dependency, as proposed by Wang and Kockelman [13]. Another feature in the MAR pattern is that there is no distinct difference in the prediction errors among the three kriging methods. In comparison to the result of MCAR, the superiority of OK is canceled out in MAR if

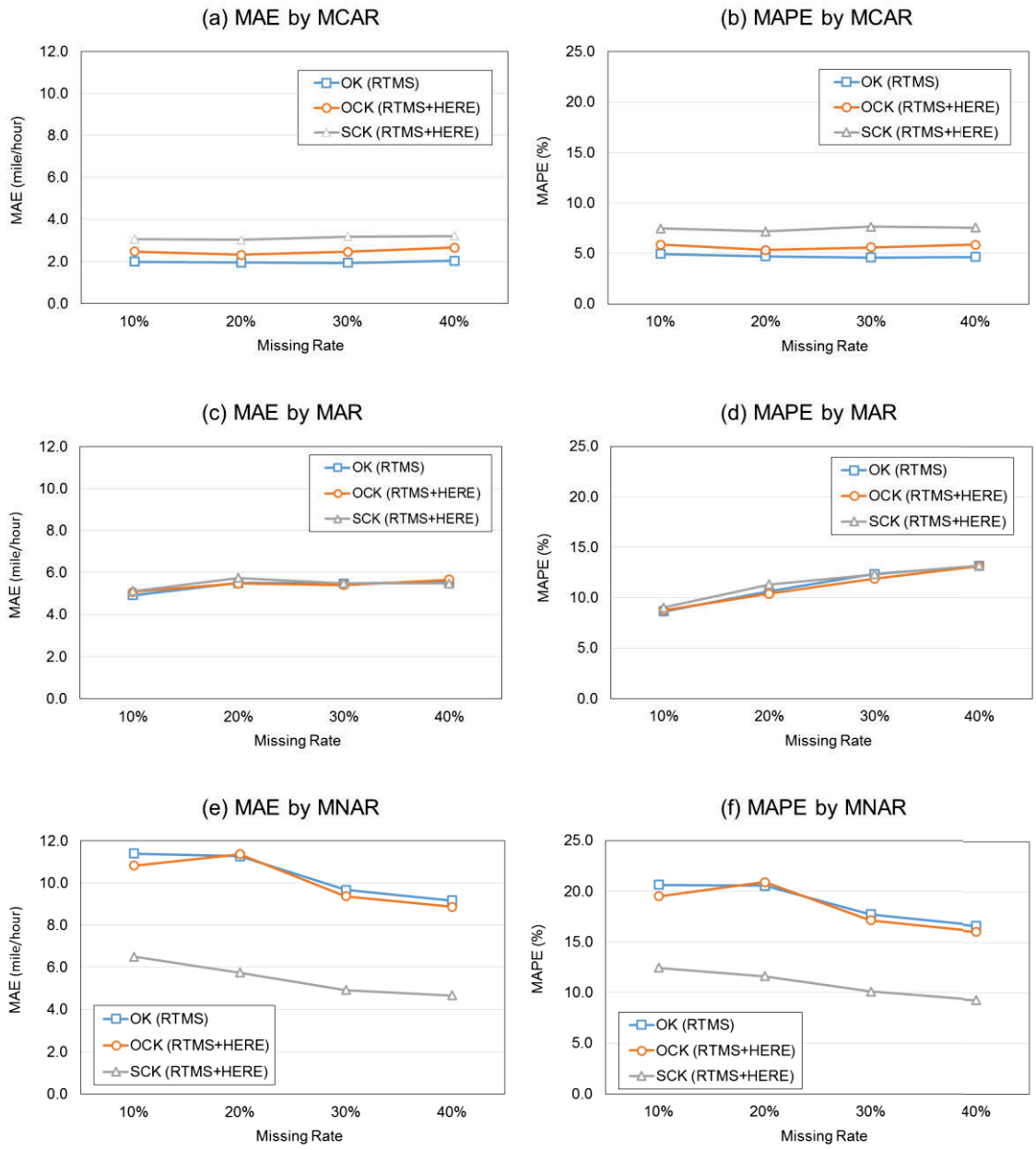


Figure 1-8 MAE and MAPE comparisons.

there are a smaller number of neighboring RTMS data points explaining the temporal dependency.

In the MNAR scenarios in Figure 1-8(e) and Figure 1-8(f), simple cokriging (SCK) outperforms the others, and the gaps in the error measures are much greater than those in the previous two missing patterns. The mean error of SCK ranges from 4.7 to 6.5 mph, while that of OK and OCK ranges from 9.4 to 11.4 mph. Likewise, the range of the mean percentage error of SCK is from 9.3% to 12.4%, while that of both ordinary kriging methods (OK and OCK) is from 16.1% to 21.0%. The prediction performances of OK and OCK are very similar in the MNAR scenarios. As discussed in the previous section, both ordinary kriging methods use an unknown local mean of a set of neighboring data points. In other words, the accuracy of the predicted value for OK and OCK could be lower than that of SCK when there are fewer or no reliable neighboring values. Since a block of data points was removed in the MNAR scenarios, the remaining neighboring data points are not as good as those in the previous two missing pattern scenarios for explaining the spatiotemporal dependency. Similar to the implication of MAR, the utility of cokriging approaches that take secondary variables into account for predicting unknown primary values is regarded as more effective when missing values are clustered over a relatively extensive spatiotemporal domain.

Given our experiments, it is clear that the prediction errors of the MNAR scenarios decrease gradually as missing rate increases. This is mainly because the proportion of high-speed observations in the validation dataset of a higher missing rate scenario increases. Since these high-speed observations are similar to the mean of the RTMS data used in this study, the overall imputation error decreases as the size of missing block increases, consequently.

Conclusion

Most of the traffic flow data imputation studies in the past have focused on investigating imputation techniques with a single data source. However, by increasing the number of data collecting sensors and other technologies over the last decade there are abundant alternative data sources that can be used to complement each other, suggesting the potential to use multi-source data to enhance imputation for missing traffic flow data. To this end, this study proposed a spatiotemporal cokriging approach to impute high resolution traffic speed data by using two complementary data sources, RTMS and HERE speed data. Two cokriging methods, ordinary and simple, were used and evaluated by comparing them with the spatiotemporal ordinary kriging method. The radar detector data (RTMS) and probe vehicle data (HERE) were used for the cokriging-based imputation approaches as primary and secondary variables, respectively.

Three different missing patterns in the spatiotemporal domain with varying missing rates were tested to evaluate the prediction performances of the cokriging methods. Generally, all kriging methods provide reliable and consistent results over various missing rate under the MCAR patterns (random in time and location) with very small, negligible errors. Because sufficient highly correlated neighboring data points exist for each missing value in the spatiotemporal context, the prediction performance is hardly influenced by missing rates. Among the three methods, ordinary kriging outperforms the others. For MAR patterns (random only in location), the difference in the performances of all methods is not prominent. For this reason, using only a primary data source for MCAR and MAR patterns can be more cost-effective than using multiple data sources. Meanwhile, one possibility of improving the prediction performance of cokriging is to consider secondary data sources if they are highly correlated with the primary data. In this study, each HERE data link covers multiple RTMS stations, implying the lower resolution of the secondary variable and relatively weak correlation between two

data sources. This may cause the cokriging methods to have the test errors. Nevertheless, it was underlined that using secondary data sources with the simple cokriging can improve prediction results when the missing pattern follows MNAR (not random in time and location). Traffic flow data collected from detectors on roads usually have missing values for a variety of reasons. Considering the fact that traffic detector data may be missing because of system malfunctions, no power supply, and maintenance, the patterns are likely to be MAR or MNAR and using spatiotemporal cokriging with multiple data sources can be beneficial for imputation.

CHAPTER II

SHORT-TERM TRAFFIC SPEED PREDICTION FOR A LARGE-SCALE ROAD NETWORK

This chapter presents a modified version of a research paper by Bumjoon Bae, and Lee D. Han.

Abstract

Short-term traffic prediction has been an essential part of real-time applications in modern transportation systems for the last few decades. Despite the recent progress in the voluminous models and data sources, many existing studies have focused on prediction for either a single or a few locations. In addition, the spatio-temporal dependency in the traffic data was narrowly accounted for. Therefore, this paper proposes a new short-term traffic speed prediction algorithm that can efficiently cope with the complexity and immensity of the prediction process derived from the network size and amount of data in order to provide accurate predictions in real time. This algorithm consists of two modules: (a) principal component analysis (PCA) for data dimensionality reduction and feature selection, and (b) multichannel singular spectral analysis (MSSA) for multivariate time-series data prediction. A large amount of traffic data is efficiently compressed by PCA with high accuracy, then used as an input in the nonparametric multivariate time-series analysis. The algorithm was compared with a vector autoregressive (VAR) model to predict traffic speeds five minutes ahead for a 21.3 mile-long highway segment, using the traffic detector data, and for 451 mile-long segment, using probe-based speed data in Tennessee. The proposed algorithm is found to provide accurate predictions with a computation time of less than one second without training. Furthermore, the proposed algorithm shows a better prediction performance under congested flow conditions, compared to VAR. This indicates that the proposed algorithm is suitable for real-time prediction and scalable for a large network analysis.

Introduction

Traffic speed is one of the fundamental variables that characterize traffic flow. It is not only a traffic performance measurement of roadway systems, but also an input for estimating other measurements such as travel time, vehicle emission, traffic noise, and so on [33]. Hence, traffic speed prediction is a core function required in modern traffic management and operation systems. In the last few decades, various short-term traffic speed prediction models and algorithms have been developed for real-time intelligent transportation systems (ITS) applications.

Although there is no absolute definition of how long the 'short-term' is, the prediction time step varies from one second to five minutes in the literature [34-42]. And the prediction horizon has been set as the range from one minute to two hours in advance through multi-step runs [43]. According to a recent comprehensive review on short-term traffic forecasting by Vlahogianni, Karlaftis [43], the majority of the previous studies used univariate models with traffic detector data at a single location on a highway. Statistical time-series models and neural network (NN) type models present a noticeable frequency of use. The time-series models include vector autoregressive (VAR) models for multivariate prediction [38, 39], spatial temporal autoregressive moving average (STARMA) for considering spatiotemporal correlation [41], generalized autoregressive conditional heteroscedasticity (GARCH) for capturing unexpected speed dynamic shifts [40], and adaptive Lasso regression for improving prediction performance by minimizing error variance [44], and so on. On the other hand, a variety of NN based models has also been proposed for speed prediction. These models are known to provide a more accurate prediction for nonlinear traffic flow compared to the classical statistics models [45-47]. Further, these models have been tested with Kalman filters or wavelet transformation technique primarily for denoising traffic data [37, 48, 49].

In the literature, the mean absolute prediction error (MAPE) of existing studies ranges from 2.5% to 15.0% for five-minute predictions [34, 35, 37, 38, 41, 44, 45, 49]. Although the effects of variability in the time step on prediction performance has not been addressed sufficiently, the prediction error shows generally a linear association with the length of a prediction time step or the number of time steps increase [34, 35, 41, 44, 50]. A few studies compared the prediction performances of congested and non-congested traffic flow conditions. They showed that the prediction errors of congested conditions are approximately three times higher than those of non-congested conditions [38, 40]. The speed threshold to define congestion varies over the studies, ranging from 30 miles per hour (mph) to 40 mph.

Despite the extensive studies on short-term traffic speed prediction, few have attempted to address the following limitations. The existing studies applied five minutes as a prediction time step without considering its effects. This is mainly because five minutes had been used most frequently in literature and the available data resolution was five minutes. Furthermore, there was insufficient information in the literature on computation time evaluation as real-time applications, which is helpful for other researchers and practitioners. In addition, many of the previous studies have been done on the short-term prediction for a single or several locations, in which a spatio-temporal dependency of traffic data was not sufficiently considered.

This paper proposes a new short-term traffic speed prediction algorithm for a large-scale road network. To support real-time and proactive traffic operations, the proposed algorithm aims to predict the future traffic flow conditions accurately and quickly without training a model. It is a nonparametric and data-adaptive algorithm that can handle a large-scale spatiotemporal speed data within a short amount of time.

The remainder of this chapter is in this manner. The next section details the methodologies used in the proposed algorithm. Then, the data sources and

different aspects of testing performance are described. Next, the prediction performance of the proposed algorithm is compared with that of vector autoregressive (VAR) model that has been successful in the past 10 years [38, 39, 43]. Finally, a discussion on the results and conclusion are drawn.

Methodology

The proposed algorithm consists of principal component analysis (PCA) and multichannel singular spectrum analysis (MSSA) (see Figure 2-1). First, PCA is used to extract features and reduce dimensions of the data. Then, MSSA is used for multivariate time-series prediction using the principal components from PCA. This approach has achieved satisfactory performance in medical image processing studies [51, 52].

Unlike the statistical prediction models such as autoregressive integrated moving average (ARIMA), MSSA, a multivariate extension of singular spectrum analysis (SSA) is a nonparametric, data-adaptive time-series analysis method that does not require any assumptions, such as stationarity of the data, linearity of the model, or normality of the residuals [53, 54]. These features make MSSA useful [53, 55-57]. Hence, SSA and MSSA have been widely applied recently in many disciplines such as economics, medical image processing, climatology research, etc. [51, 58, 59]. More theoretical and mathematical details of SSA can be found in [57] and [60]. Furthermore, using the principal components (PC) as an input of MSSA allows the prediction to be made based on spatio-temporal dependencies in the data. According to Asif, Kannan [61], PCA consistently provides high reconstruction accuracy over different compression rates for spatiotemporal traffic data.

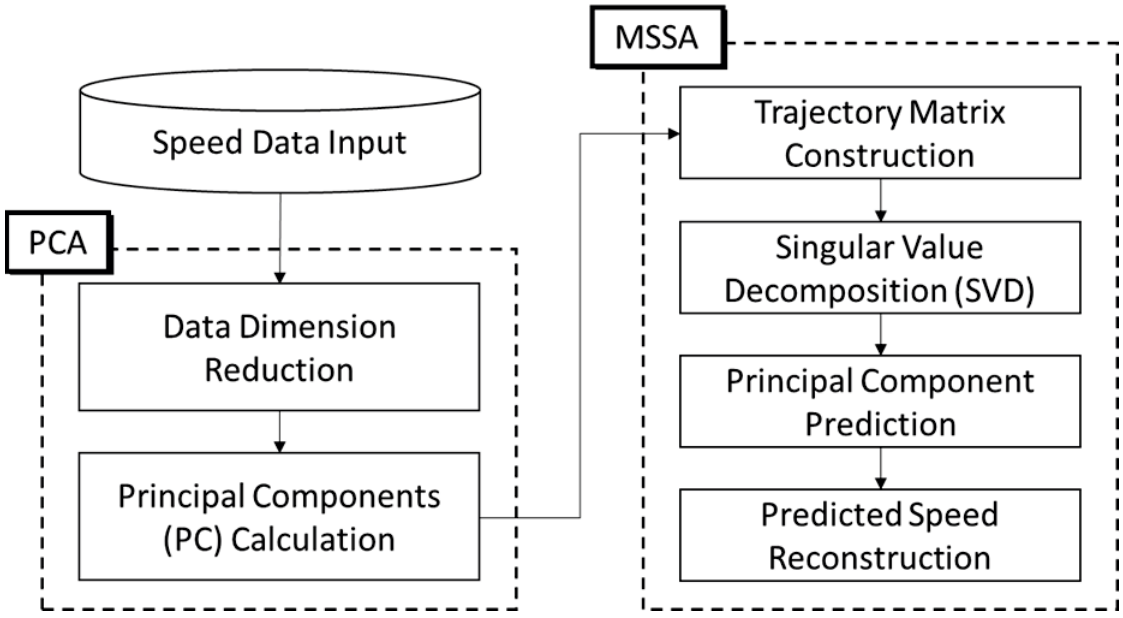


Figure 2-1 Proposed speed prediction algorithm.

Principal Component Analysis (PCA) for Data Dimension Reduction and Feature Extraction

Principal component analysis (PCA) is a widely used multivariate statistical procedure used for data dimension reduction and feature extraction [62]. It is an orthogonal transformation method that projects the original data onto the spaces of linearly uncorrelated variables where the variance is maximized based on eigenvalues and eigenvectors. Therefore, the principal components (PC), the transformed data can be used as an input for a variety of post analyses.

The speed observation x_{it}, $1 \leq i \leq n$, $1 \leq t \leq p$, with i representing location and t representing time, gives the multivariate time-series data as

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \tag{9}$$

The covariance matrix is calculated as

$$C = \frac{1}{p}\sum_{t=1}^{p} \Psi_t \Psi_t^T = \Phi\Phi^T \tag{10}$$

where, $\Psi_t = X_t - \mu$, which is the vector difference between the observations at time t and the mean of X, μ. Since $\Phi = \frac{1}{\sqrt{p}}\left[\Psi_1, \ldots, \Psi_p\right]$, the dimension of the covariance matrix C is $(n \times n)$.

As the road network size to be analyzed is increased, especially when $n \gg p$, calculating $\Phi\Phi^T$ and its eigenvectors becomes more intractable. In order to near-real-time analysis, Turk and Pentland [63] proposed to use $\Phi^T\Phi$ instead of $\Phi\Phi^T$, which reduces the dimension from $(n \times n)$ to $(p \times p)$. This approach is very common in image processing analysis where the input data at each time step is usually a 2-dimentional image. For example, if the input data size is $(n \times n)$, the size of time-series data, X is $(n^2 \times p)$, so $\Phi\Phi^T$ gives $(n^2 \times n^2)$

covariance matrix. More details about the relationship of $\Phi^T\Phi$ with $\Phi\Phi^T$ is provided in Equation (11) through Equation (14).

The eigenvector v_i is defined as

$$\Phi^T\Phi v_i = \lambda_i v_i \tag{11}$$

where, λ_i is the eigenvalue of $\Phi^T\Phi$ denoted by $\lambda_1 \geq \cdots \geq \lambda_p$. If Φ is multiplied in both sides of Equation (11),

$$\Phi\Phi^T\Phi v_i = \lambda_i \Phi v_i \tag{12}$$

and using Equation (10) and Equation (12),

$$C\Phi v_i = \lambda_i \Phi v_i. \tag{13}$$

Then, Equation (13) can be expressed as

$$Cu_i = \lambda_i u_i. \tag{14}$$

Therefore, $\Phi\Phi^T$ and $\Phi^T\Phi$ have the same eigenvalues and their eigenvectors have the relationship as $u_i = \Phi v_i$.

Finally, the orthogonally transformed data, Y is computed by using the $(p \times p)$ eigenvectors, u as follows.

$$Y = X^T u \tag{15}$$

The resultant $(p \times p)$ matrix, Y from Equation (15) is used as an input data for the following MSSA procedure.

Multichannel Singular Spectrum Analysis (MSSA) for speed prediction

The first step of MSSA is called embedding, which means mapping each univariate time series into multivariate series using subsets of the univariate time series. This procedure is similar to a time series analysis based on moving

average calculation [56]. For example, using the *k*th column of Y , $\left[y_1^{(k)}, y_2^{(k)}, \dots, y_p^{(k)}\right]^T$, the resultant matrix of embedding, called trajectory matrix, is defined as

$$y^{(k)} = \begin{bmatrix} y_M^{(k)} & y_{M+1}^{(k)} & \cdots & y_p^{(k)} \\ y_{M-1}^{(k)} & y_M^{(k)} & \cdots & y_{p-1}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ y_1^{(k)} & y_2^{(k)} & \cdots & y_{p-M+1}^{(k)} \end{bmatrix} \tag{16}$$

where, *M* is the embedding dimension (also called *window length*) which is an arbitrary integer that $2 \leq M \leq p$. Alessio [55] provides a "reasonable" range of M that is greater than the number of data points in which one oscillation to be detected and less than $p/5$. However, it is better to choose the value of *M* based on the comparison of the results from different values of *M*. Therefore, a sensitivity analysis was conducted in the case study to investigate the effects of choosing the values of *p* and *M* in the next chapter.

$\mathrm{y}_{cr}^{(k)}$ is the centered matrix of $y^{(k)}$ based on each row mean, the trajectory matrix of MSSA is made as

$$Y_{cr} = \begin{bmatrix} y_{cr}^{(1)} \\ y_{cr}^{(2)} \\ \vdots \\ y_{cr}^{(K)} \end{bmatrix} \tag{17}$$

where *K* is the number of selected PCs corresponding to the *K*th largest eigenvalues in Equation (14) ($1 \leq k \leq K$). Y_{cr} is a ($KM \times p'$) matrix and $p' = p - M + 1$. What to be estimated is the next column of Y_{cr}. This is defined as

$$Z = \left[y_{cr,p+1}^{(1)}, y_{cr,p}^{(1)}, \dots, y_{cr,p-M+2}^{(1)}, \dots, y_{cr,p+1}^{(K)}, y_{cr,p}^{(K)}, \dots, y_{cr,p-M+2}^{(K)}\right]^T. \tag{18}$$

In this study the number of PCs from Equation (15) are selected to explain over 99.7% of the total variance in order to minimize losing information of the original data, *X*. In MSSA, the row length of matrix Z gets longer as the road network size increases, compare to SSA. Then, the dimension becomes much larger after being squared in the following step. Figure 2-2 shows the percentage of data dimension reduction by using PCA for MSSA. Compare to the case of using MSSA without PCA, for example, the data dimension in MSSA is reduced by approximately 90% by PCA if the original data dimension of $(n \times p)$ is (300×100). A different number of PCs can be selected by employing information criteria, such as AIC, ICOMP, etc.

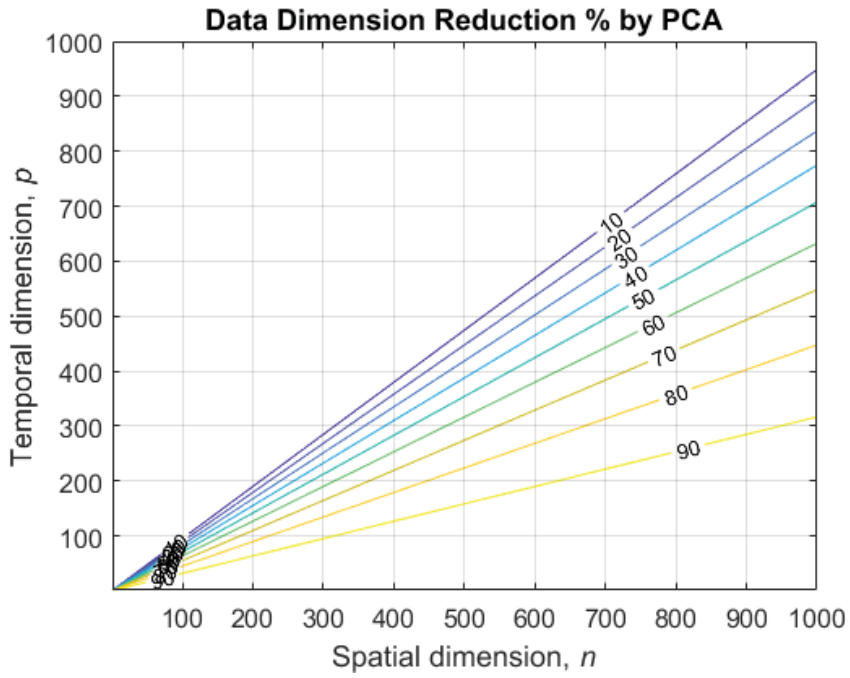


Figure 2-2 Data dimension rate of MSSA by using PCA.

The next step of MSSA is a singular value decomposition (SVD) of the squared trajectory matrix, $C_Y = Y_{cr} Y_{cr}^T$. The elements of the lagged-covariance matrix C_Y reflect the linear correlation between the all pair of patterns in the

embedding window. Thus, the recurring patterns in the time series result in a relatively high covariance in C_Y [57]. Through SVD, C_Y is decomposed into orthogonal eigenvectors as follows.

$$C_Y = E\Lambda E^T \tag{19}$$

where, E is the eigenvectors of C_Y which are the singular vectors of Y_{cr}, and Λ is a diagonal matrix that consists of ordered values, equal or greater than zero, whose square roots are the singular values of Y_{cr}. Then, the *L* largest eigenvalues from Λ and corresponding eigenvectors from E are selected for prediction as Equation (20). In this study *L*=*p* is applied which is large enough to contain the most significant eigenvectors. Through this step, the recurring patterns in the time series can be separate and the noise in the data can be removed [56].

$$W = \left[E^{(1)}, E^{(2)}, \dots, E^{(L)}\right] \tag{20}$$

Using the selected $(KM \times L)$ eigenvector matrix W, the estimation of *Z* is given as the least-squares problem as follows [51, 52, 60].

$$\text{minimize } (Z - WW^T Z)^2 \tag{21}$$

This implies that the evolution of the next vector in the trajectory matrix follows the same law of the other adjacent vectors [64].

Then, Z can be decomposed as,

$$Z = RP + Q \tag{22}$$

where $P = \left[y_{cr,p+1}^{(1)}, y_{cr,p+1}^{(2)}, \dots, y_{cr,p+1}^{(\mathrm{K})}\right]^T$. The $(KM \times K)$ and $(KM \times 1)$ restriction matrices, *R* and *Q* are defined as follows.

$$R = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & 1 & \cdots & \vdots \\ \vdots & 0 & \cdots & \vdots \\ \vdots & \vdots & \cdots & 1 \\ \vdots & \vdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix}, Q = \left[0, y_{cr,p}^{(1)}, \dots, y_{cr,p-M+2}^{(1)}, \dots, 0, y_{cr,p}^{(K)}, \dots, y_{cr,p-M+2}^{(K)}\right]^T \quad (23)$$

By decomposing Equation (21) with Equation (22), the future component of the time series data can be obtained as Equation (24) [51, 52]

$$P = (I - R^T W W^T R)^{-1} R^T W W^T Q \quad (24)$$

where, *I* is a $(K \times K)$ identity matrix.

Finally, the predicted speed is calculated by re-centering the values of *P* and multiplying them with the eigenvectors from Equation (14).

Case Study

Data description

The proposed prediction algorithm was applied to speed data for Interstate 40 (I-40) in Tennessee from two data sources: (a) traffic detector data, named Remote Traffic Microwave Sensors (RTMS), which is collected every 30 seconds from over 1,000 traffic detector stations on interstate highways in Tennessee, and (b) probe-based link speed data, named National Performance Management Research Data Set (NPMRDS). For RTMS, the detector stations are located only in major urban areas of the state. Therefore, 41 stations in the 21.3 mile-long westbound I-40 segment were selected, which is a major corridor in Knoxville, Tennessee. The stations are on average 0.5 miles from each other. Traffic speeds for the intermediate locations in 0.1-mile increments between two

consecutive stations were interpolated using the adaptive smoothing method [65] in order to augment the spatial resolution of the data by 213.

The speed data from September 23 and September 30 in 2016, both of which were Fridays, were collected from the detectors and averaged in five minutes, i.e., the data dimension is (213×288) for each day. Both days were selected based on the fact that there was no incident in the first day while there was a severe incident on the second day. The incident was verified by the traffic incident data log from the local transportation management center (TMC). Since prediction of unexpected events, such as crashes, adverse weather conditions, etc., in the spatiotemporal domain is highly intractable, it is worth testing how quickly the speed prediction algorithm can adapt or how sensitive it is to sudden changes in traffic conditions.

In order to evaluate the proposed algorithm performance for a longer road segment, i.e., larger data dimension, the NPMRDS data were used. For NPMRDS, the spatial coverage is the entire interstate highway systems in the state. In this study, the five-minute average speeds of NPMRDS for the 298 road links of a 451-mile-long I-40 westbound segment on February 3rd, 2017 were collected. Please note that five minutes are the highest resolution for the available NPMRDS dataset, i.e., the data dimension is (298×288). Figure 2-3 shows examples of the data visualizations.

Performance measures

To evaluate the prediction performance of the proposed algorithm, three error measures were used, which are the mean absolute error (MAE) and mean absolute percentage error (MAPE). They are defined as follows.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}_i| \tag{16}$$

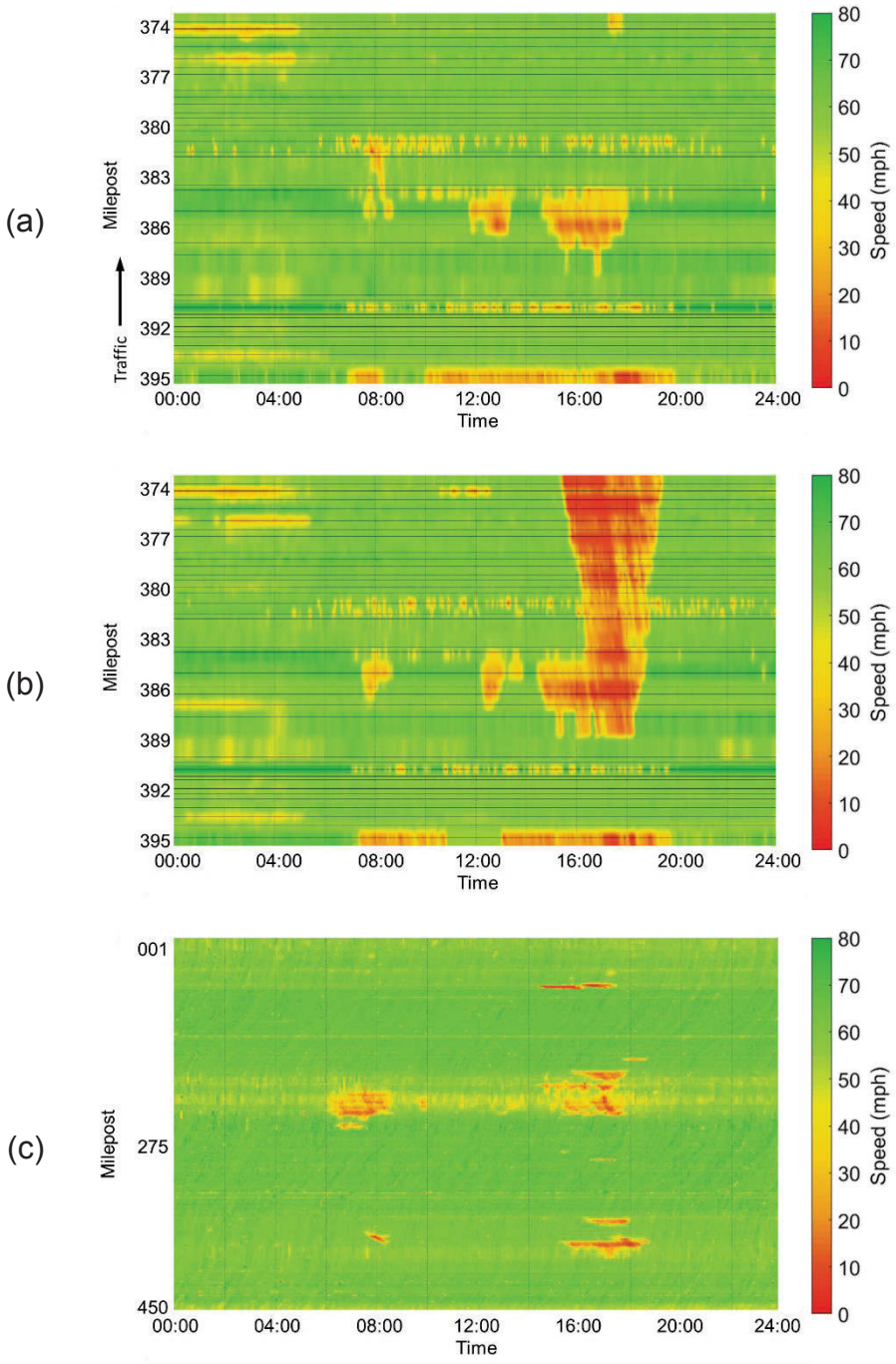


Figure 2-3 Speed data visualizations: (a) RTMS – September 23, 2016; (b) RTMS – September 30, 2016; and (c) NPMRDS – February 3, 2017.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|x_i - \hat{x}_i|}{x_i} \times 100 \quad (17)$$

where, x_i is the observed traffic speed and $\hat{x}_i$ it the predicted traffic speed.

Data resolution selection

To choose an optimal prediction interval is an important issue which depends on the type of ITS applications, algorithms and data sources [43]. In order to investigate the effect of the data resolution on the short-term traffic speed prediction, a sensitivity analysis framework was applied. The need for a sensitivity analysis is mainly due to the nonparametric characteristic of PCA-MSSA, i.e., it does not allow to test statistical significance of parameter estimates. Four datasets of the 24-hour traffic speeds from RTMS were generated by different aggregation levels: 0.5-, 1-, 2.5-, and 5-minute and used in a preliminary analysis. To make predictions for the target time in the future, the iterative predictions are made, i.e., the predicted values are added to the initial data for the next prediction. Table 2-1 shows the average prediction performance for 5-minute prediction. Each prediction was made using the past thirty data points. To predict the next five-minute traffic speed, for example, the prediction process is implemented ten times iteratively using the 30-second dataset. As the number of prediction steps increases, the prediction error increases. This is because the error in the current prediction is transferred to the next prediction step. Therefore, five minutes gave the lowest errors for the five-minute prediction. The following analyses were made using the data aggregated in five minutes.

Input Data Dimension and Window Length Selection

The effects of choosing different data length p and window length M were investigated in a sensitivity analysis. Here the range of 0.5-6 hours for both p and M was considered using the 5-minute RTMS data of September 23 and

Table 2-1 Temporal scale effects on 5-minute prediction performance using RTMS.

MOEs	Data resolution (Number of prediction steps)			
	0.5 min (10)	1 min (5)	2.5 min (2)	5 min (1)
MAE (mph)	3.40	3.31	3.12	3.03
MAPE (%)	9.67	9.16	8.23	7.94

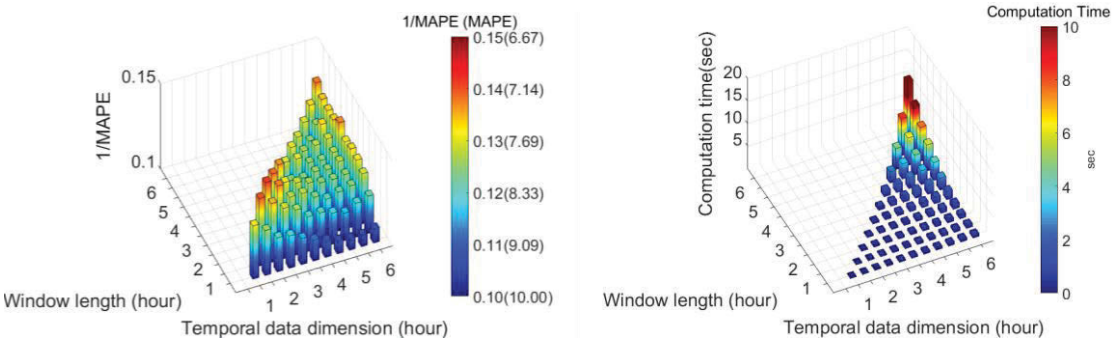
September 30 in 2016 and NPMRDS data on February 3, 2017. In order to choose proper values of p and M, MAPE and computation time for one-step prediction were compared as shown in Figure 2-4 and Figure 2-5. Please note that the vertical axis of Figure 2-4(a) and Figure 2-5(a) represent 1/MAPE for better recognition of the best result. Figure 2-4(a) shows that there is a gradual increase in MAPE with increase of both of p and M in the range of 1-5.5 hours. The computation time in Figure 2-4(b) also shows the same pattern; however, it increases much more rapidly as p and M get closer to six hours. Similar patterns were observed in Figure 2-5. Based on these sensitivity results, $p = 18$ (1.5 hours) and $M = 12$ (1 hours) for RTMS – September 23, 2016, $p = 24$ (2 hours) and $M = 18$ (1.5 hours) for RTMS – September 30, 2016, and $p = 36$ (3 hours) and $M = 18$ (1.5 hours) for NPMRDS were applied.

Prediction Performance

Parametric versus Nonparametric Methods

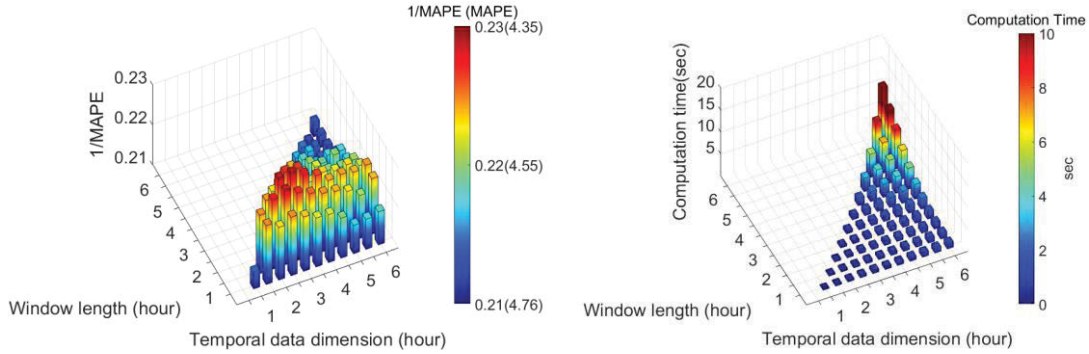
To evaluate the proposed algorithm, the speed prediction results for the next five minutes were compared to those of a parametric model, VAR(*k*). In this study, the order of the model *k* was determined to be within the range of 1-8 (i.e., $k = 1, \dots, 8$) based on the goodness of fit of the model using Akaike's information criterion (AIC) [66]. The 24-hour historical speed data were used for each prediction target time point to train the VAR(*k*) model. The RTMS dataset was used to make 288 predictions for September 23 and 30, 2016. In order to compare the computation time, both methods were implemented on the same platform with Intel® Core™ i7 processor (3.60GHz) with 8GB memory.

In this study, restricted VAR models were used. Unrestricted VAR models using a full covariance matrix for parameter estimation is not suitable for real-time data analysis on a large-scale network for these reasons: First, the model estimation time is too long because a large number of parameters will be estimated. For example, an unrestricted VAR(1) model with *n* = 213 has 68,373



(a) (b)

Figure 2-4 RTMS with different temporal dimension and window length (September 30, 2016): (a) MAPE and (b) computation time.



(a) (b)

Figure 2-5 NPMRDS with different temporal dimension and window length: (a) MAPE and (b) computation time.

$(= n + nAR \cdot n^2 + n(n+1)/2$, where the number of the autoregressive matrix, $nAR = 1$) parameters to be estimated, while a restricted model has only 639 ($= 3n$). Therefore, estimating an unrestricted VAR model takes too long when either the network *n* or the autoregressive lag *k* is large. Second, the residual process of the unrestricted model is likely to have a non-positive definite covariance matrix which makes parameter estimation impossible.

In order to investigate the effect of applying PCA in the proposed algorithm, MSSA without PCA, referred to hereafter as MSSA, was also tested. In addition, based on the fact that it is more likely to use a pre-trained parametric model in practice, the VAR(*k*) model was separated into two types: (a) a model whose parameter estimates are updated for each prediction, denoted as On-VAR(*k*); and (b) a model whose parameter values are fixed once the model is trained priorly, denoted as Off-VAR(*k*). Please note that the model order *k* of On-VAR(*k*) is not updated for each prediction step; otherwise, training a model takes an excessive amount of time, making short-term prediction harder to achieve. Therefore, the same order k of Off-VAR(*k*) was applied to On-VAR(*k*). For the same reason, the On-VAR(k) model was trained using five-hour historical data for each prediction target time.

Table 2-2 summarizes the 5-minute prediction performances of these four methods. For Scenario 1 non-incident condition, Off-VAR(7) outperforms the others. In this scenario, traffic flow is very stable in terms of speed except for the congestion around milepost 386 during afternoon peak hours. For such cases, the speed data hold high stationarity and the parametric model fits the data well. The error level of PCA-MSSA is slightly higher than both VAR models and MSSA. In comparison with Off-VAR(7), as depicted in Figure 2-6(a), the level of error of PCA-MSSA is slightly higher than that of Off-VAR(7) across the overall error range. This may result from the information loss of data in the dimension reduction procedure or the misspecified length of the input data and embedding window.

Table 2-2 Comparison of 5-minute prediction performance for RTMS.

		PCA-MSSA	MSSA	On-VAR(k)	Off-VAR(k)
Scenario 1 (No incident)	Model selection	$p = 18$ $M = 12$	$p = 18$ $M = 12$	$k = 7$	$k = 7$
	MAE (mph)	2.31	2.26	2.19	1.96
	MAPE (%)	4.98	4.90	4.76	4.26
Scenario 2 (Incident)	Model selection	$p = 24$ $M = 18$	$p = 24$ $M = 18$	$k = 8$	$k = 8$
	MAE (mph)	2.46	2.39	2.52	2.41
	MAPE (%)	6.56	6.40	7.02	7.26
Average Computation time (sec)		0.05	6.78	114.20	0.22

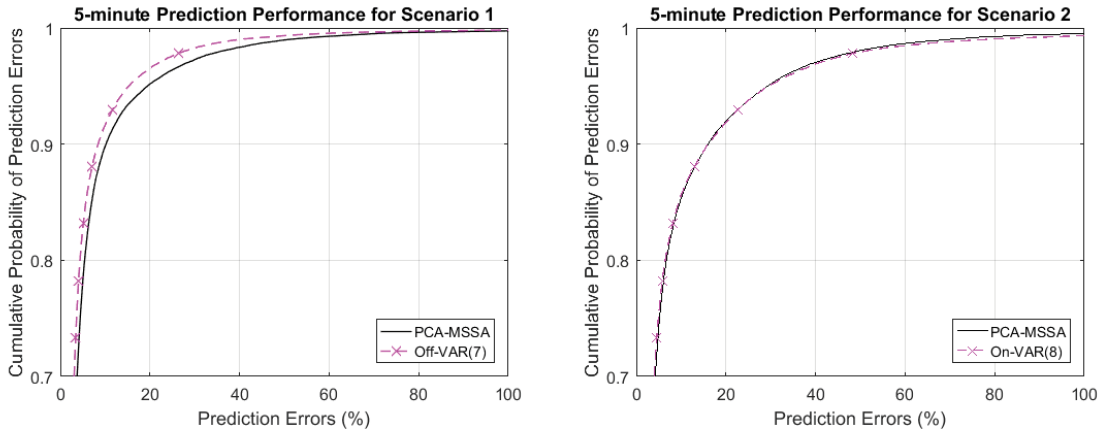
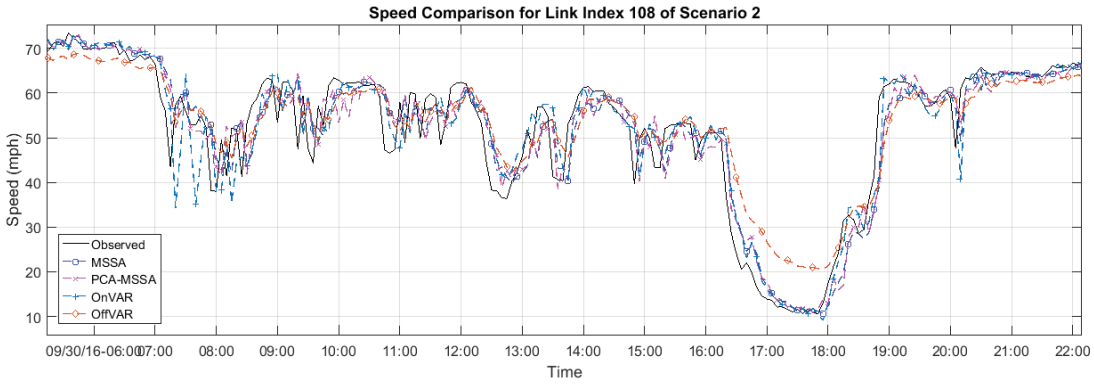


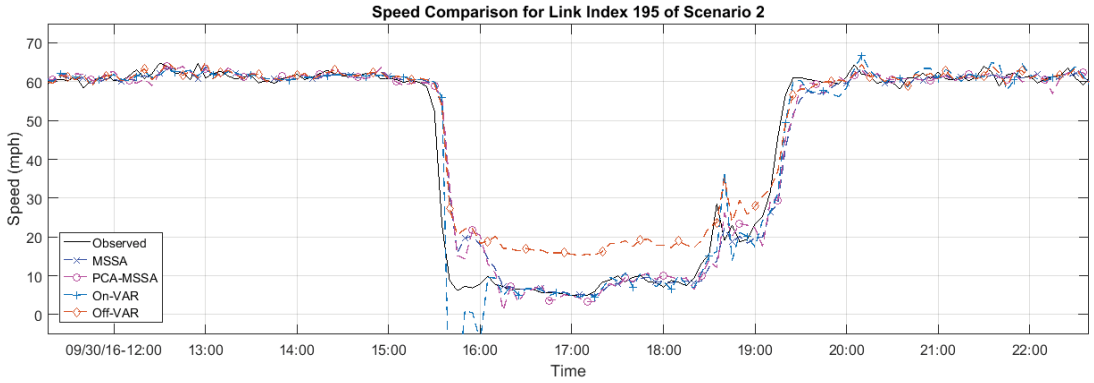
Figure 2-6 Prediction performance of PCA-MSSA and VAR.

Despite the different prediction performance in Scenario 1, traffic prediction for a free-flow condition is not challenging. In other words, prediction of traffic conditions during the transitions to and from congested flow over time and space should be paid attention more. The traffic condition in Scenario 2 shows such instability in the speed data caused by a severe incident. As shown in Table 2-2, MSSA and PCA-MSSA outperform both VAR models. The MAPE of 6.40% from MSSA is slightly better than 6.56% from PCA-MSSA. Since the same dimension of input data was employed, it is probable that the different performance was caused by PCA. Contrary to the result in Scenario 1, On-VAR outperformed Off-VAR in Scenario 2 in terms of MAPE. On-VAR model predicts the congested flow better than Off-VAR by updating parameter estimates for each prediction. In order to evaluate the performance of PCA-MSSA for congested traffic flow, its prediction error range is compared with that of On-VAR in Figure 2-6(b). Although the cumulative probability error curves of both methods are very similar, they intersect at around 25%. This indicates that the average error level of PCA-MSSA is relatively lower for low speed conditions, compared to On-VAR.

Figure 2-7 shows the predicted speed profiles of four methods at selected locations. The Figure 2-7(a) location is in a weaving section where two major interstate highways are merged. Recurrent afternoon congestion was intensified due to an incident that occurred downstream around 3:00 to 4:00 PM. All the predicted profiles, except for Off-VAR, show similar patterns and follow the observed speed fluctuation. However, On-VAR tends to produce overfitted results when the traffic state changes from free-flow to congestion in the morning peak hours. The same pattern of On-VAR is also present in Figure 2-7(b). The average performance measurement in space is shown in Figure 2-8. Both PCA-MSSA and MSSA outperform the VAR models during the congested time period. With the emergence of congested traffic flow, all the error measures are increased. However, both MSSA algorithms quickly adapt to the changes of flow



(a)



(b)

Figure 2-7 Predicted speed profiles: (a) location index 108 and (b) location index 195.

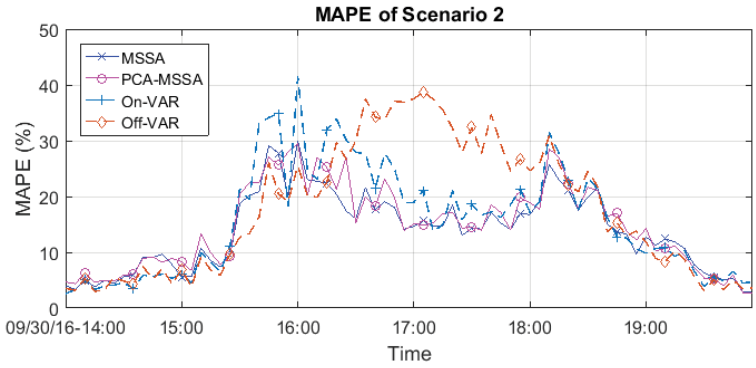


Figure 2-8 Prediction errors during an incident event.

states so that their error measures are decreased. Although the error of On-VAR also decreases as the model adapts to the congested state, the error level is high when the traffic state transition begins.

Numerical accuracy of the prediction is obviously important in the model comparison. However, comparing different models based solely on the accuracy may be not fair, since other factors such as computation time, required data size, the level of expertise, etc., are important as well [43, 67]. This is true because the purpose of the proposed method focuses on the near-real-time traffic speed prediction for a large road network. Therefore, the computation time to make a one-step prediction with the four methods was compared. The computation time of PCA-MSSA was considerably shorter than those of MSSA and On-VAR model. PCA-MSSA took only 0.05 seconds to predict traffic speed 5 minutes ahead for the 213 different locations; the MSSA algorithm without PCA took 6.78 seconds on average. Although Off-VAR also processed the data quickly, i.e., 0.22 seconds on average, the 0.5-hour training time is not accounted for. In practice, however, the model training time should be considered because periodical updates of parameter values may be needed to retain or enhance the current performance. Combined with the comparison result of prediction accuracy in Scenario 2, the computation efficiency of PCA- MSSA shows that the proposed algorithm is more suitable than the others to predict traffic speed for a large-scale network in real time. Because of the data dimensionality reduction feature, the proposed method is scalable for a larger road network analysis.

Multi-Step Speed Prediction

The prediction error is accumulated as the number of prediction steps increases. In order to test the prediction performance of PCA-MSSA for the future in longer than five minutes, predictions were made for up to 30 minutes ahead and compared with Off-VAR. Table 2-3 summarizes the multi-step speed prediction results. Over the multiple prediction steps, the average error of PCA-MSSA

Table 2-3 Prediction performance for multi-step predictions.

		MOEs	5 min ahead	10 min ahead	15 min ahead	20 min ahead	25 min ahead	30 min ahead
Prediction steps			1	2	3	4	5	6
Scenario 1 (No incident)	PCA-MSSA	MAE (mph)	2.31	2.57	2.75	2.89	3.01	3.11
		MAPE (%)	4.98	5.57	6.01	6.35	6.64	6.88
	Off-VAR(7)	MAE (mph)	1.96	2.22	2.41	2.55	2.67	2.78
		MAPE (%)	4.26	4.81	5.22	5.55	5.85	6.11
Scenario 2 (Incident)	PCA-MSSA	MAE (mph)	2.46	2.86	3.14	3.37	3.57	3.75
		MAPE (%)	6.56	7.86	8.85	9.66	10.38	11.04
	Off-VAR(8)	MAE (mph)	2.41	2.87	3.24	3.55	3.84	4.10
		MAPE (%)	7.26	9.13	10.76	12.24	13.58	14.81

showed a moderate increase from 4.98% to 6.88% in Scenario 1. In contrast, the more rapid increase of error from 6.56% to 11.04% was observed in Scenario 2. As a reference, the prediction performance measures of Off-VAR are provided together. As the comparison result for the single-step prediction, the errors of Off-VAR are slightly lower than those of PCA-MSSA in multi-step prediction, while the opposite comparison results present in Scenario 2.

It is difficult to directly compare the prediction performance of the proposed algorithm with the results reported in the literature due to different data sources, times, and locations with different study designs. Despite this reason, such comparison may help researchers gain a general sense of the current state in speed prediction studies. The error level of the proposed algorithm is slightly lower or comparable to that of NN-based and parametric time-series models in the literature [34, 37, 40, 44, 50].

Algorithm Scalability Investigation

To test the scalability of the PCA-MSSA algorithm for speed prediction, NPMRDS data were used in this study. The obtained data covers the entire westbound I-40 segment in Tennessee. The data dimension is (298×288) i.e., one day of 5-minute speeds from 298 road links. The majority of the speeds in the dataset represent the free-flow condition except for those of major urban areas during peak hours. Therefore, the computation time is the major interest in this comparison, although the error measures are also presented in Table 2-4. The comparison result of the computation time is very similar to that in the RTMS case, despite the NPMRDS data dimension being almost 40% larger than the RTMS dataset. PCA-MSSA took 0.36 seconds for one-step prediction, while MSSA took 7.24 seconds. The computation time of Off-VAR is smallest in the comparison. However, the model estimation time of 2.5 hours is not reflected in the result.

Table 2-4 Comparison of 5-minute prediction performance for NPMRDS.

	PCA-MSSA	MSSA	On-VAR(k)	Off-VAR(k)
Model selection	$p = 36$ $M = 18$	$p = 36$ $M = 18$	$k = 3$	$k = 3$
MAE (mph)	2.29	2.39	2.24	2.26
MAPE (%)	4.32	4.53	4.16	4.54
Average Computation time (sec)	0.36	7.24	80.34	0.49

Conclusion

Previous short-term traffic prediction studies have investigated a vast number of models and algorithms in the last two decades. Nevertheless, there is still room to progress prediction performance by employing data-driven multivariate models and corresponding large datasets for real-time traffic controls and operations. This paper proposed a short-term traffic speed prediction algorithm to cope efficiently with the complexity and immensity of the prediction process derived from the network size and amount of data. The proposed algorithm, named PCA-MSSA, consists of two techniques: (a) principal component analysis (PCA) for data dimensionality reduction and (b) multichannel singular spectral analysis (MSSA) for multivariate time-series data prediction.

The prediction performance of PCA-MSSA was compared to the parametric time-series model, vector autoregressive (VAR). For the incident scenario, PCA-MSSA outperformed VAR and it provided speed predictions in near-real-time. Although the pre-trained VAR model showed slightly lower prediction errors on average for the non-incident scenario, PCA-MSSA still predicted the speed with comparable accuracy levels. This is mainly because PCA-MSSA uses the compressed spatiotemporal traffic data as an input and it is a nonparametric data-adaptive method. In contrast, VAR is a more complex model that requires more data, and it estimates a tremendous number of parameters for a large-scale network analysis. This result shows that PCA-MSSA is suitable for real-time traffic speed prediction and scalable for a large network analysis. To identify the effect of PCA in the proposed algorithm, the results were compared to the case of MSSA without PCA. Interestingly, a trade-off between the accuracy and computation time was reported. Using PCA can reduce computation time significantly with a relatively small compromise in prediction accuracy.

Further research should be directed at the following challenges: (a) improving the prediction accuracy of the proposed algorithm during non-recurring events through cooperation with automatic incident detection algorithms and more advanced PCA methods; (b) adding a self-learning process after the predicted values are validated; (c) developing a dynamic optimization process to select the length of historical data and embedding window length of the algorithm over time; and (d) predicting travel time based on the predicted speed and conducting comparative evaluations.

CHAPTER III
SPATIO-TEMPORAL TRAFFIC QUEUE DETECTION FOR HIGHWAYS

Abstract

When traffic demand exceeds capacity because of demand fluctuations, crashes, work zones, and special events, a traffic queue is formed on a highway. Traffic queues cause potentially hazardous situations at the end of the queue where drivers unexpectedly face slowed or stopped traffic while approaching at high speed. Therefore, detecting and predicting a queue is vital for protecting it. This study presents a real-time spatio-temporal traffic queue detection algorithm that builds on traffic flow fundamentals combined with a statistical pattern recognition procedure. Using flow-density data, traffic flow phase is classified as either congested or uncongested flow in a probabilistic manner, based on Gaussian mixture models for each location in such a way that detects the traffic phase transitions. Next, empirical shock wave speeds of the detected queue between downstream and upstream locations are calculated in a time-space domain, which will predict the queue arrival time at the next upstream detecting location. The proposed detection algorithm was applied to detect traffic queues using traffic detector data from Interstate 40 in Knoxville, Tennessee. The detection results show that the algorithm detects queues successfully by accounting for varying queueing conditions and different queue types.

Introduction

Monitoring and predicting the evolution of traffic queues in a spatio-temporal domain are the most necessary tasks to prevent primary and secondary crashes on highways. A physical shock wave is generated at the end of a queue when a traffic flow changes from one condition to another, e.g., from uncongested flow to congested flow. Then the shockwave propagates either upstream or downstream at a different speed, depending on the differences in traffic conditions (i.e.,

densities and flow rates) between the upstream and downstream of the end of a queue. In the case of upstream propagation, the upstream vehicles approaching at high speed may encounter the shockwave without enough response time, increasing the probability of a traffic crash. Therefore, if an advisory message is transmitted to the upstream vehicles based on the predicted information of queue propagation, the shockwave can be absorbed and weakened, thereby stabilizing traffic flow.

The term “queue” has been defined in various ways in literature. Highway Capacity Manual 2010 [68] defines a queue as “a line of vehicles waiting to be served” in a system and a queued state as “a condition when a vehicle has slowed to less than 5 mph”. Stephanopoulos, Michalopoulos [69] defines queue length for an intersection as “the length of the roadway section behind the stop line where traffic conditions range from the capacity to jammed density” in a flow-density diagram. In spite of the different and insufficient queue definitions for freeway facilities in the literature, a common condition is that a queue is formed when the system demand exceeds its capacity [33]. It is difficult to measure the traffic demand directly from traffic flow data when the flow is at or near capacity at a bottleneck. However, One can infer the presence of the excessive demand if high densities and low speeds are observed upstream of the bottleneck [33]. In traffic flow theory, a breakdown is the transition from uncongested to congested flow and observed as a speed drop occurring with queue formation [68]. Therefore, the spatio-temporal evolution of a queue can be identified by detecting the phase transitions based on the data patterns of the fundamental traffic variables at multiple locations in real time.

This study proposes a short-term traffic queue detection and prediction algorithm that is adaptive to local traffic conditions for detecting phase transitions, i.e., transition from uncongested to congested flow and vice versa, and trace the propagation of congestion in real-time. In order to detect the transition, a Gaussian mixture model (GMM) based classification algorithm was developed to

fit the data distributions of the congested and uncongested flows of each traffic detector station on a highway. GMM is a probability density estimation method that uses a mixture of multivariate Gaussian distributions to fit a distribution of given data. The advantage of mixture models including GMM is that analysts can control the number of components, i.e., control the trade-off between the computational efficiency of parametric methods and model fitting flexibility of non-parametric methods. For parameter estimation of GMM, the expectation maximization (EM) algorithm is used [70].

The next section presents a literature review on traffic queue detection and describes the data used in this study. The following section explains the proposed algorithm and related methodologies. The last two sections present the result of a case study that applied the queue detection algorithm using the detector data; conclusions comprise the final section.

Literature Review

Previous studies on traffic queues have focused on estimating queue length for interrupted flow facilities such as signalized intersections. Since it is important to manage queue lengths for intersections, queue lengths are used to measure traffic signal performance and optimize signal timing plans. Many queue-related studies for uninterrupted facilities, meanwhile, have focused on estimating queue delay and queue length for a work zone. For the methodology aspect, the literature can be classified in two major categories: (a) cumulative traffic input-output approach and (b) traffic shock-wave approach.

Input-Output Approach

Queue length is a function of traffic demand and capacity. The initial model, proposed by Webster [71], calculated the time of the queue dissipation and effective queue size by input-output analysis. After the start-up lost time from the

onset of a green signal, the queued vehicles are discharged at saturation flow, and after the onset of a red signal, another queue forms based on an assumed arrival rate. Since the effective queue size is defined as the number of vehicles in the queue waiting for service at an instant in jam density [72], a constant average density throughout cycles is assumed in the range between the jammed flow and capacity flow. However, it has been pointed out that density is time varying within a cycle and the assumption of constant average density can lead to miscalculation of the effective queue size [72]. Sharma, Bullock [73] evaluate two input-output models. One is a simple model in which only advance detector is used to track vehicle arrivals, and another model uses advance and stop bar detectors to utilize the headway information. The root mean squared error of both models was shown as less than 0.15 vehicle for average maximum queue length by evaluation with field data. These models cannot estimate queue lengths or produce inaccurate estimation results when queue rear exceeds beyond the detector because arriving vehicles cannot be detected [74].

Deterministic queueing analysis has been used to estimate queueing delay and queue lengths, in which vehicle arrival and service distributions are specified as deterministic distributions. Cassidy and Han [75] proposed vehicle delay and queue length estimation methods for two-lane highways. Deterministic queueing theory was applied to compute queue lengths. Jiang and Adeli [76] proposed a queue delay and queue length estimation algorithm for freeway work zones. The estimation is made based on the estimated work zone capacity. The queue length is estimated by a deterministic macroscopic queueing model.

Shock Wave Approach

Lighthill and Whitham [77] and Richards [78] explained traffic flow phenomena on the basis of shock wave theory, using a theoretical fundamental diagram called LWR theory. In their model, the flow rate is assumed as a function of the vehicle density [77]. Although the shock wave theory is derived from the conservation

law of vehicle counts, which accounts for traffic flows going into and out from a roadway segment, the queue length estimation models in this type use the shock wave speed directly. Geroliminis and Skabardonis [79] proposed an analytical models for predicting platoon arrival profiles and queue length along signalized arterials. They employed a Markov decision process to model traffic dispersion behaviors between successive signal intersections, and then shock wave speeds based on the LWR theory were used to estimate queue lengths. The difference of predicted queue length between the model and simulated results was less than four vehicles. Liu, Wu [74] proposed a real-time queue length estimation method for congested signalized intersections using event-based signal and vehicle detection data. They applied LWR shockwave theory to identify break points where traffic flow states change at a loop detector location. Then, the maximum queue length can be estimated at the intersecting point of a discharge and departure shockwave speed.

However, there is some criticism of the LWR theory. Kerner [80] claimed that the LWR theory cannot explain some empirical traffic flow phenomena including a probabilistic speed breakdown occurring spontaneously at a bottleneck due to an internal local disturbance in traffic flow (i.e., transition from free flow to synchronized flow).

Location Based Data Approach

During the last decade, new attempts to estimate queue lengths in real-time are using location information of probe vehicles in a queue. Comert and Cetin [81] proposed a conditional probability model to estimate the expected queue length and its variance. Based on the assumption that the marginal probability distribution of queue length is known and the vehicle arrivals follow Poisson distribution, they found that the location information of the last probe vehicle in a queue is sufficient for estimating queue length regardless of the market penetration of probe vehicles. However, the finding is limited since it is based on

a priori knowledge of the marginal distribution and derived for undersaturated conditions. Ban, Hao [82] estimated the maximum and minimum queue lengths by detecting critical pattern changes of intersection travel times or delays based on the GPS log information.

Although these studies can also be classified as the shockwave approach, using the location information of individual vehicles can be distinguished from earlier studies where fixed location sensor data were used.

Implications

Even though the previous studies have mostly focused on the estimation of queue lengths for a signalized intersection with a single link, these estimation approaches can be employed for uninterrupted traffic flows. Traffic queues occur at signalized intersections, and they also occur because of traffic incidents and natural bottlenecks on freeways. If there are fixed, successive locational traffic sensors such as loop detectors in a study area, the input-output approach can be applied. A growing queue can be detected over time by using multiple detectors at the upstream locations. The shock wave theory can also be employed in the same sense. By capturing shock wave speeds for successive detector stations, the locations of the queue ends can be estimated collectively. If individual vehicle trajectory data are available in real time for a highway where the detectors are deployed, the estimation result can be improved or validated.

The expected challenges for each approach are the following:

- Input-Output: Since multiple highway links—a link here is defined as the roadway segment between two detector stations—should be considered for detecting a queue, calculating accurate inflow and outflow traffic volumes in a subject road segment would be difficult due to on- off-ramp flows and limitations on spatial coverage and temporal resolution of detector data, e.g., 30 seconds.

- Shock Wave: This method uses the relationship of traffic flow q and density k in traffic flow fundamentals. In general, the q-k relationship is estimated linearly as a concave line so that shockwave speed is calculated by selecting two single points on the line. This cannot reflect the variance or probabilistic phenomena in the q-k relationship, especially for a congested flow. In addition, unlike the signalized intersection case where one of both traffic states for a shockwave speed is the jam density, traffic incident or bottleneck may not be connected to a complete stop of the flow or complete blockage of all lanes. The error in a shock wave speed estimate may produce significantly inaccurate queue length.
- Location-Based Information: This type of data is usually unavailable, particularly for a real-time traffic application.

Methodology

Proposed queue detection algorithm

Using traffic detector data collected from each detector station, traffic flow phase is identified as either a congested or uncongested flow over time, based on the station's unique flow-density pattern in the previous days. Then, congestion is detected in the flow-density domain by using the phase identification results collectively for multiple stations along a highway. Finally, by connecting the onset of congestion at each station, shock wave speeds are calculated and the queue arrival time at the next upstream station is predicted (Figure 3-1).

Overall, the proposed queue detection algorithm consists of these steps:

- Traffic flow phase identification: For each station, traffic flow phase is initially identified as one of the following classes: 'uncongested,' 'transitional,' and 'congested' flow based on speed (see Figure 3-2(a)). In this study, 45 mph and 15 mph are determined to obtain the minimum samples for each flow

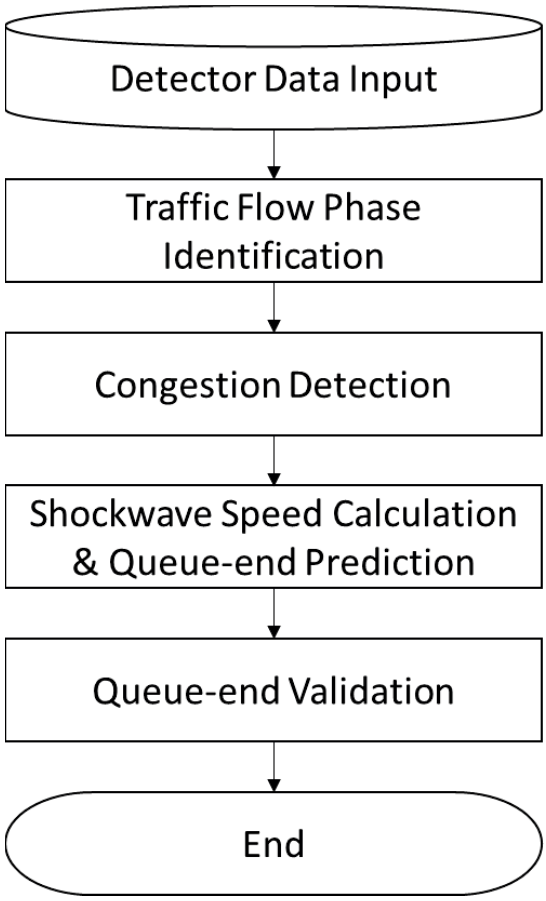
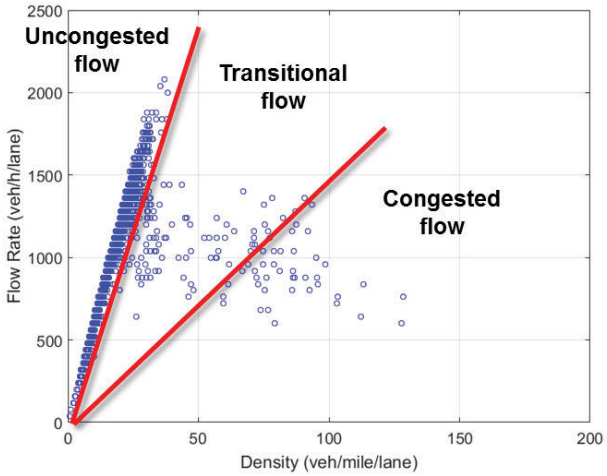
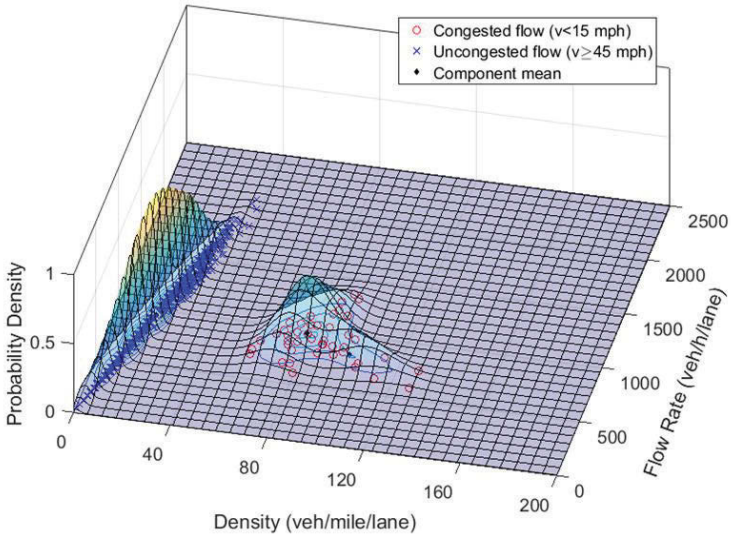


Figure 3-1 Proposed queue detection algorithm.



(a)



(b)

Figure 3-2 Phase identification: (a) three phases in a flow-density plot and (b) an example of estimated data distributions using GMM.

as thresholds for the initial phase identification (i.e., 0-15 mph: congested flow, 15-45 mph: transitional flow, 45+ mph: uncongested flow). The distributions of congested flow and uncongested flow are estimated in a flow-density diagram using GMMs as shown in Figure 3-2(b). Then, each new input data point is classified by comparing the likelihood of both phase classes.

- Traffic congestion detection: The phase information identified for each station in the previous step is used collectively to detect congestion occurrence at multiple locations and times (see Figure 3-3(b)).
- Shock wave speed calculation and queue arrival time prediction: In order to calculate shock wave speeds, the data points on the boundary of congestion in the time-space domain should be identified. For this, an even number of phase changes within a two minute time window is filtered out (see Figure 3-3(c)). Then, using the remaining boundary points where a traffic flow phase transition occurs, shockwave speed is calculated in real-time. Then, the arrival time of a queue at the next station is predicted based on the shockwave speed.
- Queue arrival time validation: The predicted queue arrival time at the next upstream station is validated by using an error measurement.

Gaussian mixture model

Let $\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}$ denote a random sample with the size of *n*, where $\boldsymbol{x_j}$ is a *p*-dimensional random vector with probability density function (pdf) of a Gaussian distribution, $f(\boldsymbol{x_j})$. For a univariate random variable x, the pdf is

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{18}$$

where, $-\infty < x < \infty$, $-\infty < \mu < \infty$, and $\sigma^2 > 0$.

For the *p*-dimensional normal density function is

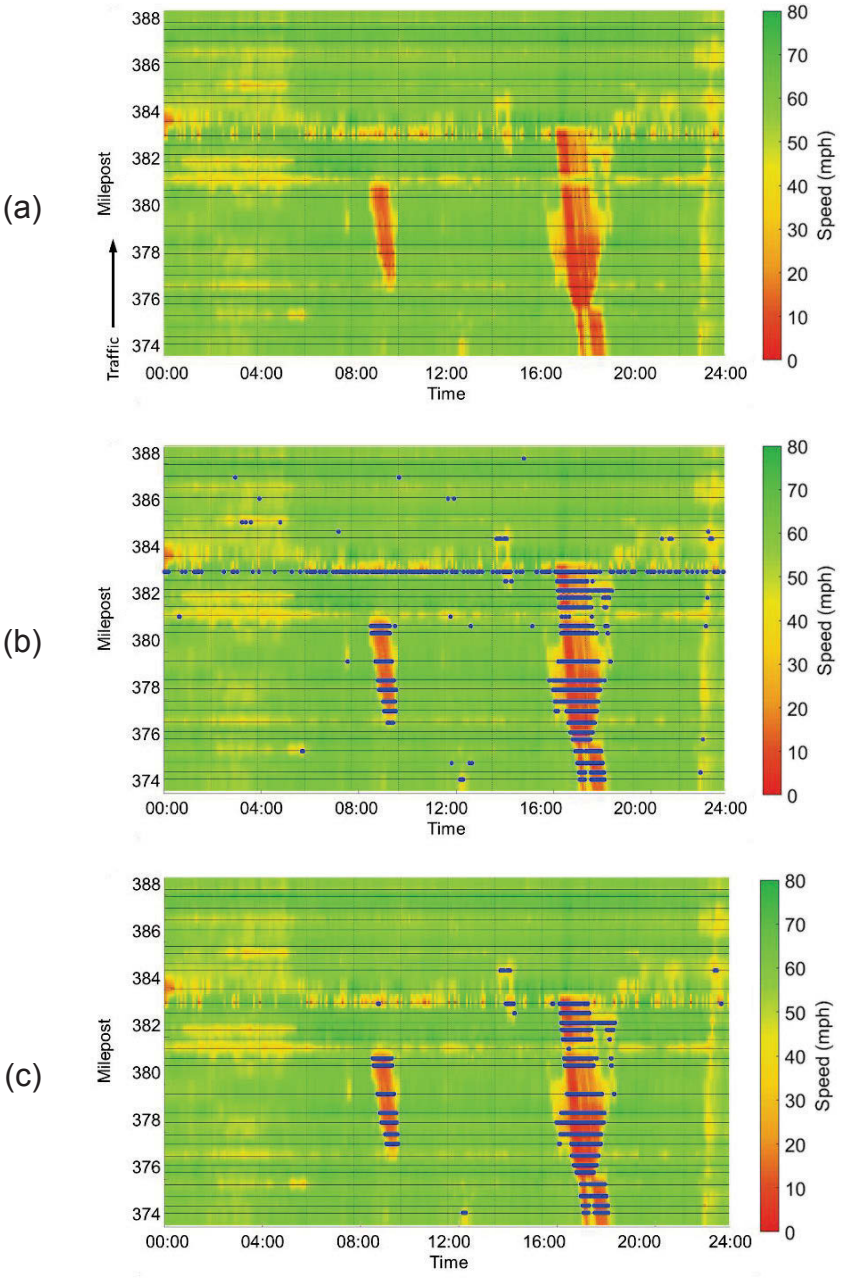


Figure 3-3 An example of congestion detection (I-40 EB on August 4th, 2016): (a) speed heat map, (b) congestion detection without filtering, and (c) congestion detection with filtering.

$$f(\boldsymbol{x_j}|\mu,\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x_j}-\mu)^T\Sigma^{-1}(\boldsymbol{x_j}-\mu)} \quad (19)$$

where, Σ is a covariance matrix that is positive definite, i.e., $\Sigma > 0$.
The probability density function of data can be represented as a Gaussian mixture distribution, which is a linear combination of *K* Gaussian distributions (or components) with the set of parameters $\boldsymbol{\Theta} = \{\alpha_{i=1\dots K}, \boldsymbol{\theta}_{i=1\dots K}\}$, for each as follows.

$$f(\boldsymbol{x_j}|\boldsymbol{\Theta}) = \sum_{i=1}^{K} \alpha_i f(\boldsymbol{x_j}|\boldsymbol{\theta}_i) \quad (20)$$

where, $f(\boldsymbol{x_j}|\boldsymbol{\theta}_i)$ is the Gaussian distribution with the *i*th parameter set $\boldsymbol{\theta_i} = \{\mu_{i=1\dots K}, \Sigma_{i=1\dots K}\}$ and π_i is the mixture weight of *i*th component which is nonnegative and sum to one, that is

$$0 \le \alpha_i \le 1 \ \ (i = 1, \dots, K)$$

and

$$\sum_{i=1}^{K} \alpha_i = 1.$$

The log likelihood for $\boldsymbol{\Theta}$ is

$$\log L(\boldsymbol{\Theta}) = \sum_{j=1}^{N} \log f(\boldsymbol{x_j}|\boldsymbol{\Theta}) = \sum_{j=1}^{N} \log \left\{ \sum_{i=1}^{K} \alpha_i f(\boldsymbol{x_j}|\boldsymbol{\theta}_i) \right\}. \quad (21)$$

It is known that there is no closed form of the maximum likelihood estimation (MLE) for $\boldsymbol{\Theta}$ of the Gaussian mixture distribution. Therefore, the Expectation Maximization (EM) algorithm is frequently used to get the parameter estimates in GMM where the MLE is computed iteratively [70].
The EM algorithm consists of two steps, E for expectation and M for maximization.

E-step: Let $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{z})$ denote the complete data vector which consists of the observed data $\boldsymbol{x}$ and its posterior probability membership variable of the *K* components $\boldsymbol{z} = \{\boldsymbol{z_1}, \dots, \boldsymbol{z_K}\}$, where each $\boldsymbol{z_i}$ is an *N*-length vector $[z_{i1}, \dots, z_{iN}]^T$. The complete-data log likelihood for $\boldsymbol{\Theta}$ is

$$\log L(\boldsymbol{\Theta}|\boldsymbol{y}) = \sum_{i=1}^{K}\sum_{j=1}^{N} z_{ij}\log\{\alpha_i f(\boldsymbol{x}_j|\boldsymbol{\theta}_i)\} \tag{22}$$

where

$$z_{ij} = P(i|\boldsymbol{x}_j, \boldsymbol{\Theta}) = \frac{\alpha_i f(\boldsymbol{x}_j|\boldsymbol{\theta}_i)}{\sum_{l=1}^{K}\alpha_l f(\boldsymbol{x}_j|\boldsymbol{\theta}_l)} \quad , \quad \text{for } i \in 1, \dots, K, j \in 1, \dots, N. \tag{23}$$

Then, the conditional expectation of the log likelihood of the complete data $\boldsymbol{y}$ given the parameter estimate on (*t*)th iteration can be written as

$$\mathrm{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) = E[\log L(\boldsymbol{\Theta}^{(t)}|\boldsymbol{y})]. \tag{24}$$

M-step: The parameter set of the (*t+1*)th iteration is determined based on the estimated z_{ij}. The mixture weights would be given simply as

$$\alpha_i^{(t+1)} = \frac{1}{N}\sum_{j=1}^{N} z_{ij}\,, \quad \text{for } i \in 1, \dots, K. \tag{25}$$

$\boldsymbol{\theta}_i^{(t+1)}$ that maximizes $\mathrm{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)})$ can be found from $\frac{\partial \mathrm{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)})}{\partial \boldsymbol{\theta}_i} = \boldsymbol{0}$ and the new mean and covariance matrix are

$$\boldsymbol{\mu}_i^{(t+1)} = \frac{\sum_{j=1}^{N} z_{ij}\boldsymbol{x}_j}{\sum_{j=1}^{N} z_{ij}} \tag{26}$$

and

$$\Sigma_i^{(t+1)} = \frac{\sum_{j=1}^{N} z_{ij} \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i^{(t+1)}\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i^{(t+1)}\right)^T}{\sum_{j=1}^{N} z_{ij}}. \quad (27)$$

The E- and M-steps are repeated until either the difference $\log L\left(\boldsymbol{\Theta}^{(t+1)}\right) - \log L\left(\boldsymbol{\Theta}^{(t)}\right)$ becomes smaller than a convergence value or the number of iteration reaches the preselected maximum value. The convergence value of 0.000001 and the maximum iteration of 1000 were used in this study.

The initial parameter values were selected by using the *k*-means clustering algorithm, where the mixture probability and covariance matrix across *k* clusters were assumed to be identical, then the centroid of each cluster is computed based on the Mahalanobis distance.

Model selection using information complexity criterion: ICOMP

Choosing the number of components *K* in the context of mixture model clustering analysis is one of the common and difficult problems in all clustering techniques [83]. Akaike's information criterion (AIC) [66] and Bayesian information criterion (BIC) [84] have been frequently used in such model selection problems. AIC evaluates the lack of fit of a model with respect to a given data, penalizing it based on the number of parameters in the model as a measure of complexity. BIC accounts for the sample size, as well as the number of parameters. However, AIC and BIC are known to be inconsistent in the mixture context. AIC tends to overestimate the number of components, and BIC tends to underestimate it [85]. Bozdogan [83] proposed the informational complexity (ICOMP) criterion of an approximate inverse Fisher information matrix (IFIM) for selecting the number of components in the mixture model with consideration of not only the lack of fit but also the model complexity. The complexity in ICOMP is not the number of parameters in the model or the sample size, but the degree of interdependence among the components of the model [83]. A model with minimum ICOMP is the best model. ICOMP with IFIM is defined as

$$ICOMP(IFIM) = -2\log L\left(\widehat{\boldsymbol{\Theta}}\right) + C_1 - C_2, \tag{28}$$

where

$$C_1 = d\log\left[\frac{1}{d}\sum_{i=1}^{K}\left\{\frac{1}{\hat{\alpha}_i}tr\left(\hat{\Sigma}_i\right) + \frac{1}{2}tr\left(\hat{\Sigma}_i{}^2\right) + \frac{1}{2}tr\left(\hat{\Sigma}_i\right)^2 + \sum_{v=1}^{p}\left(\hat{\Sigma}_i\right)_{vv}^2\right\}\right]$$

$$C_2 = (p+2)\sum_{i=1}^{K}\log\left|\hat{\Sigma}_i\right| - p\sum_{i=1}^{K}\log(n\hat{\alpha}_i) + Kp\log(2n)$$

and

$$d = Kp + \frac{1}{2}Kp(p+1). \tag{29}$$

Traffic flow identification

Once the GMMs of the congested and uncongested traffic phases are estimated for each station, new data points fed into the algorithm are classified into either phase by comparison of likelihoods. Based on the Equation (21),

$$\begin{cases}\text{phase} = congested, & \text{if } \log L\left(\boldsymbol{\Theta}_{congested}|\boldsymbol{x}^{new}\right) > \log L\left(\boldsymbol{\Theta}_{uncongested}|\boldsymbol{x}^{new}\right)\\ \text{phase} = uncongested, & \text{otherwise.}\end{cases} \tag{30}$$

where, $\boldsymbol{x}^{new}$ is the new data vector of flow and density, $\boldsymbol{\Theta}_{congested}$ and $\boldsymbol{\Theta}_{uncongested}$ are the sets of parameters of the congested flow's and uncongested flow's mixture models, respectively.

Shock wave speed calculation

In traffic flow theory, a shock wave refers to boundary conditions in a time-space domain that represents a discontinuity in flow- density states [33]. Based on the

well-known traffic flow theory of *flow=speed×density*, the shock wave speed between two states is defined as the change in flow divided by the change in density as follows.

$$\omega_{AB} = \frac{q_A - q_B}{k_A - k_B} \tag{31}$$

where, *A* and *B* denote different traffic flow-density states, ω_{AB} is shock wave speed when a state changes from *A* to *B*, q is flow, and k is density.

However, applying Equation (31) is not suitable for tracing a queue in the time-space domain in real-time. Unlike the theoretical concave curve or triangular shape in a flow-density diagram, real traffic flow-density data plots often show a reversed lambda shape and very chaotic movements on the right-hand (queued) side (see Figure 3-4) [86, 87]. Therefore, shock wave speeds calculated from real data are too sensitive for the purpose of this study. In addition, the speeds from Equation (31) represents shock waves at a given station as depicted in Figure 3-5(a), not a link between stations.

Therefore, in this study, shock wave speeds are calculated empirically between a pair of stations along a highway. Two different approaches were tested as shown in Figure 3-5(b) and Figure 3-5(c). The first approach is to use the arrival time differences between two neighboring stations. The second approach is to use the arrival time difference between the first downstream station where a queue starts to form and each upstream station that the queue reaches. The shock wave speeds from the first approach can have a greater variation, while the second reduces the variation while a queue is propagating upstream. These shock wave speeds are used to predict the queue arrival time at the next upstream station.

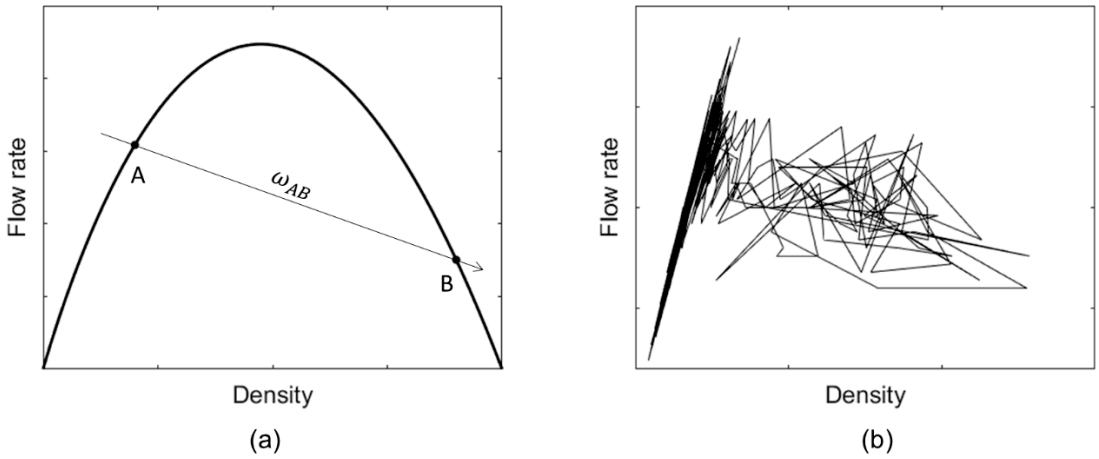


Figure 3-4 Flow-density relationship: (a) theoretical flow-density curve and shock wave speed and (b) real traffic data (station at 374.2 mile EB on August 4, 2016, 4-9 PM).

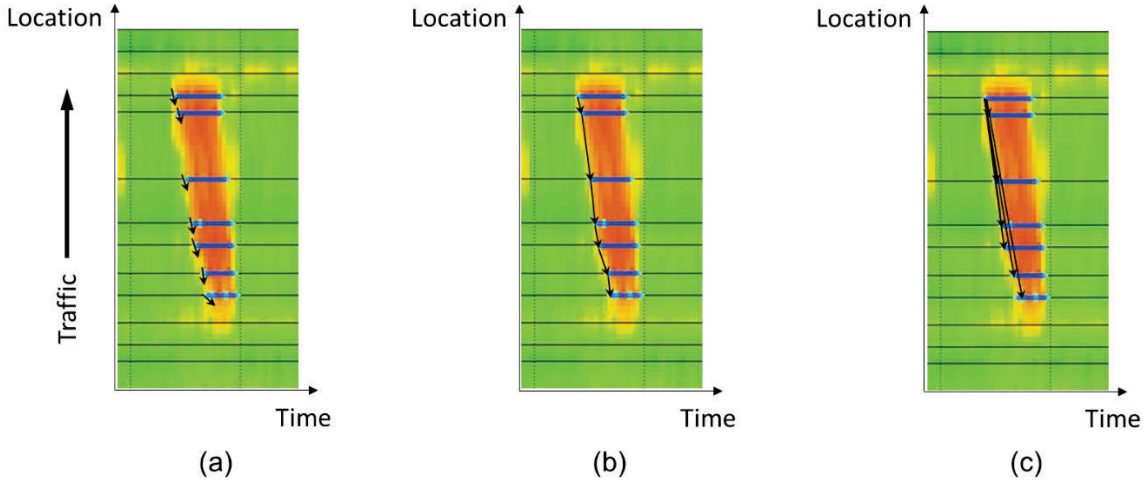


Figure 3-5 Shock wave speed calculation: (a) at each station, (b) between two neighboring stations, and (c) between the first downstream station and each upstream station.

Case Study

Data description

The traffic detector data, named the Remote Traffic Microwave Sensor (RTMS), collected in Tennessee were used in this study. Specifically, the data of Interstate 40 (I-40) in the Knoxville urban area were collected from July 11, 2016, through September 1, 2016. There are 87 detector stations on the 20.5 mile-long segment of I-40, ranging from mile marker 374.2 (west end) to 394.7 (east end), for both westbound (WB) and eastbound (EB) directions (see Figure 3-6). Due to the lack of detector stations on I-40 around the downtown area of Knoxville and the fact that no congestion is observed usually, 14 stations on I-40 EB close to the east end were excluded from this study. Therefore, the data of 73 stations on I-40 were used to implement and evaluate the proposed algorithm.

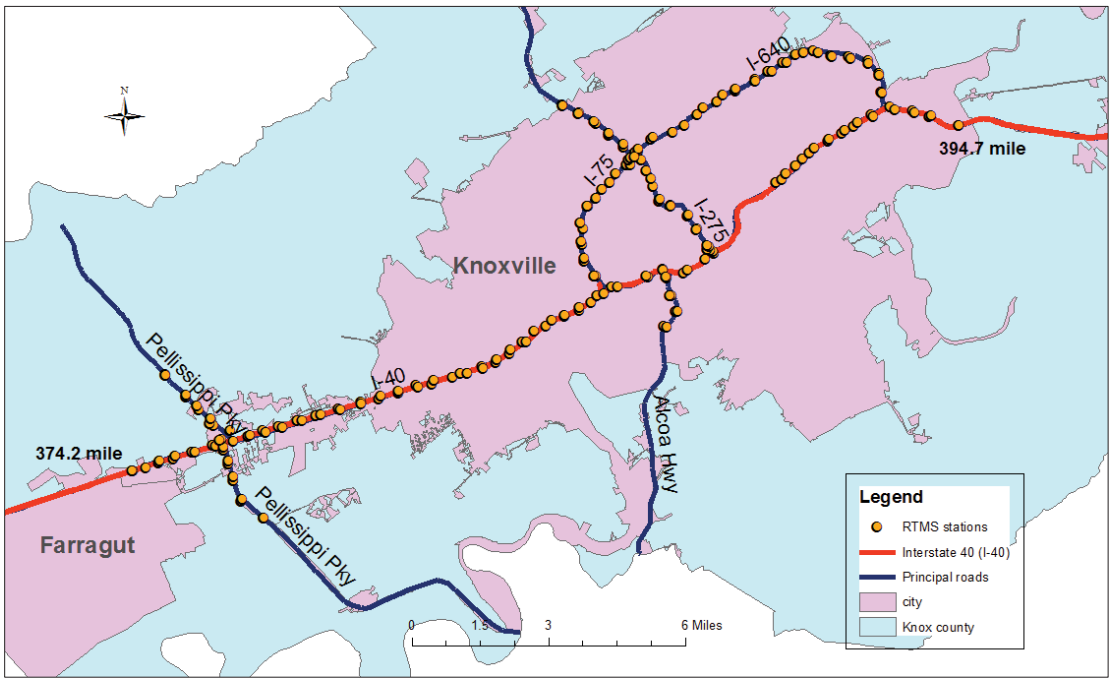


Figure 3-6 RTMS stations in Knoxville TN.

The RTMS data contains 30-second aggregated traffic count, speed, and occupancy for each lane of each station. Since the purpose of this study is to propose a “real-time” queue detection algorithm, no further temporal aggregation was made despite the unnecessary noise in the 30-second data. The lane-by-lane data were aggregated for each station.

In this study, seven days in the period between July 11, 2016 and September 1, 2016 were selected to test the proposed algorithm in which distinct queues were observed (see Figure 3-7). For each test day, all the historical data of its previous days in August 2016 were used to estimate GMMs. Due to insufficient samples for the test day of August 4, the additional data from July 11 – July 31 were used as well.

Traffic flow phase identification and congestion detection

Congestion patterns appear differently for each location due to varying capacity and demand in the time-space domain. Implementing the phase identification process independently for each location is, therefore, important for adapting to the varying traffic conditions so that the proposed algorithm can identify the phase transitions and collectively detect congestion along highways, based on the queueing condition, i.e., capacity < demand. During the identification process, the number of components of each mixture model was selected with a range of 1-6 based on ICOMP; their average number was 3.7. As mentioned in the methodology section, AIC selected more components on average, which is 4.1.

Figure 3-8 and Figure 3-9 shows the congestion detection results for the seven days before and after filtering. The speed heat maps behind the detection layers were generated by using the RTMS data with the adaptive smoothing method [65, 88]. In comparison with the speed heat map for each test day, the proposed algorithm detects the most of the low-speed traffic conditions.

Since the proposed algorithm does not directly use a fixed speed value as a threshold to detect congested flow, each congestion case shows different

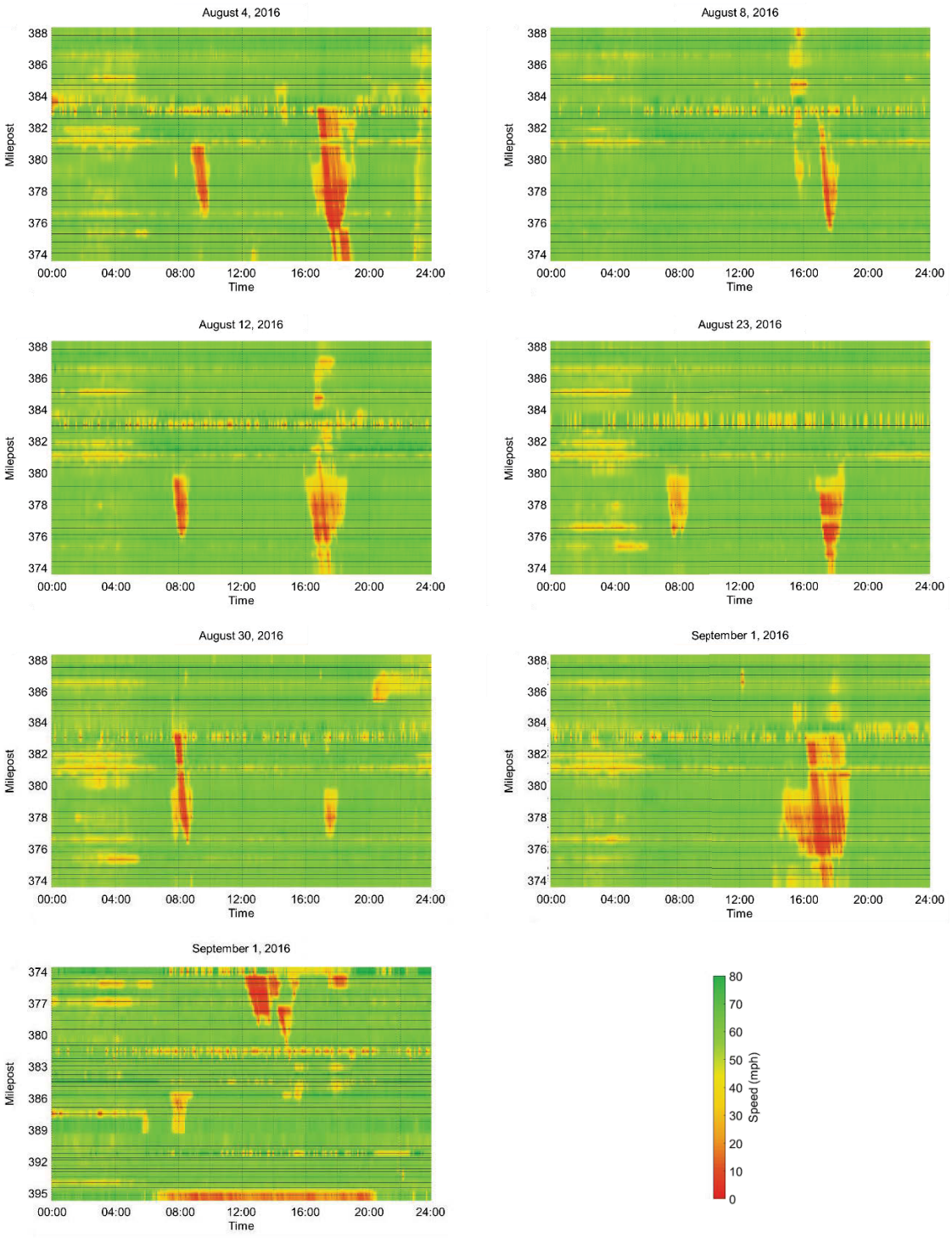


Figure 3-7 RTMS speed visualizations for the selected test days.

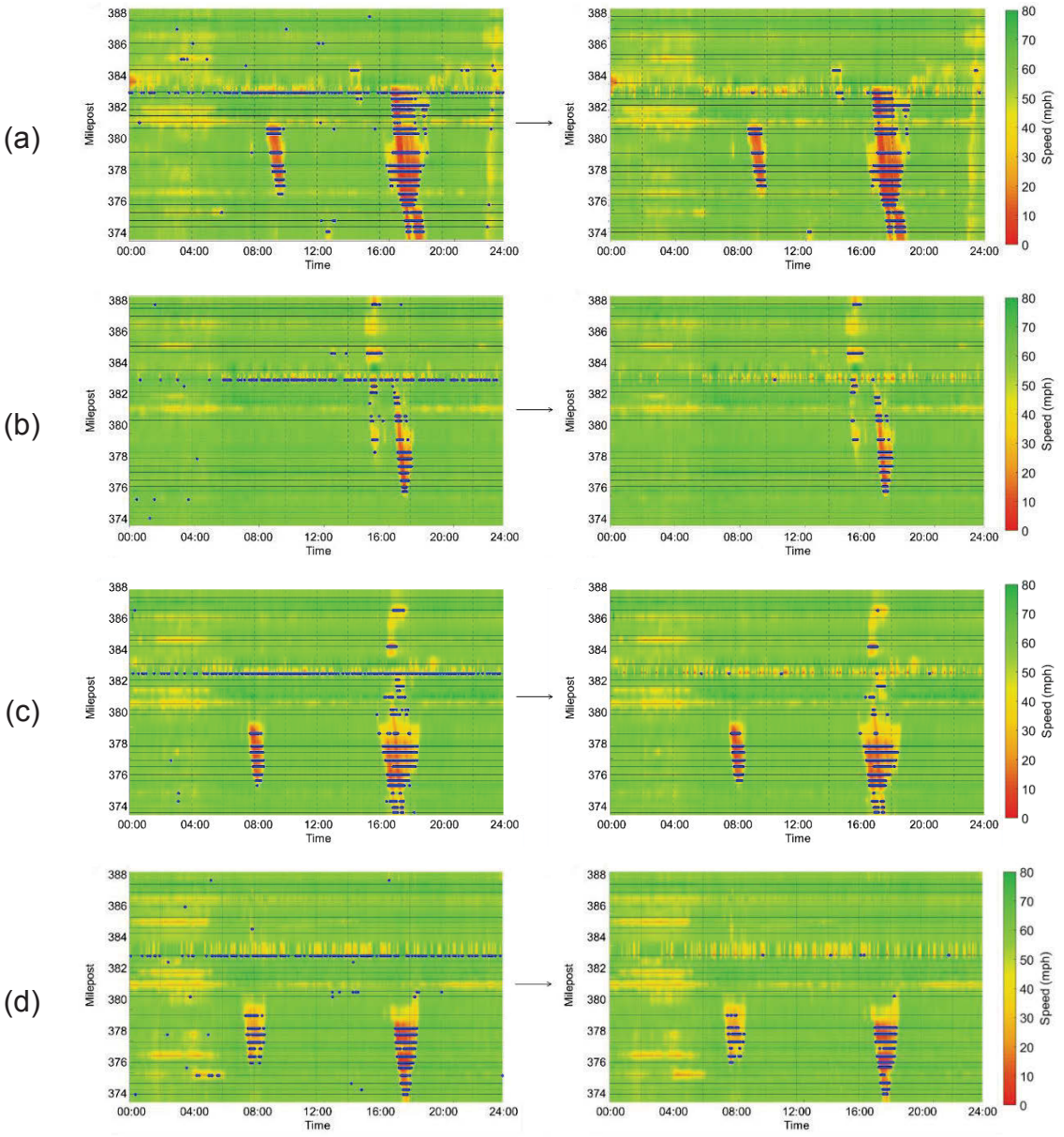


Figure 3-8 Phase identification and congestion detection results with speed heat map: (a) August 4, 2016, (b) August 8, 2016, (c) August 12, 2016, and (d) August 23, 2016.

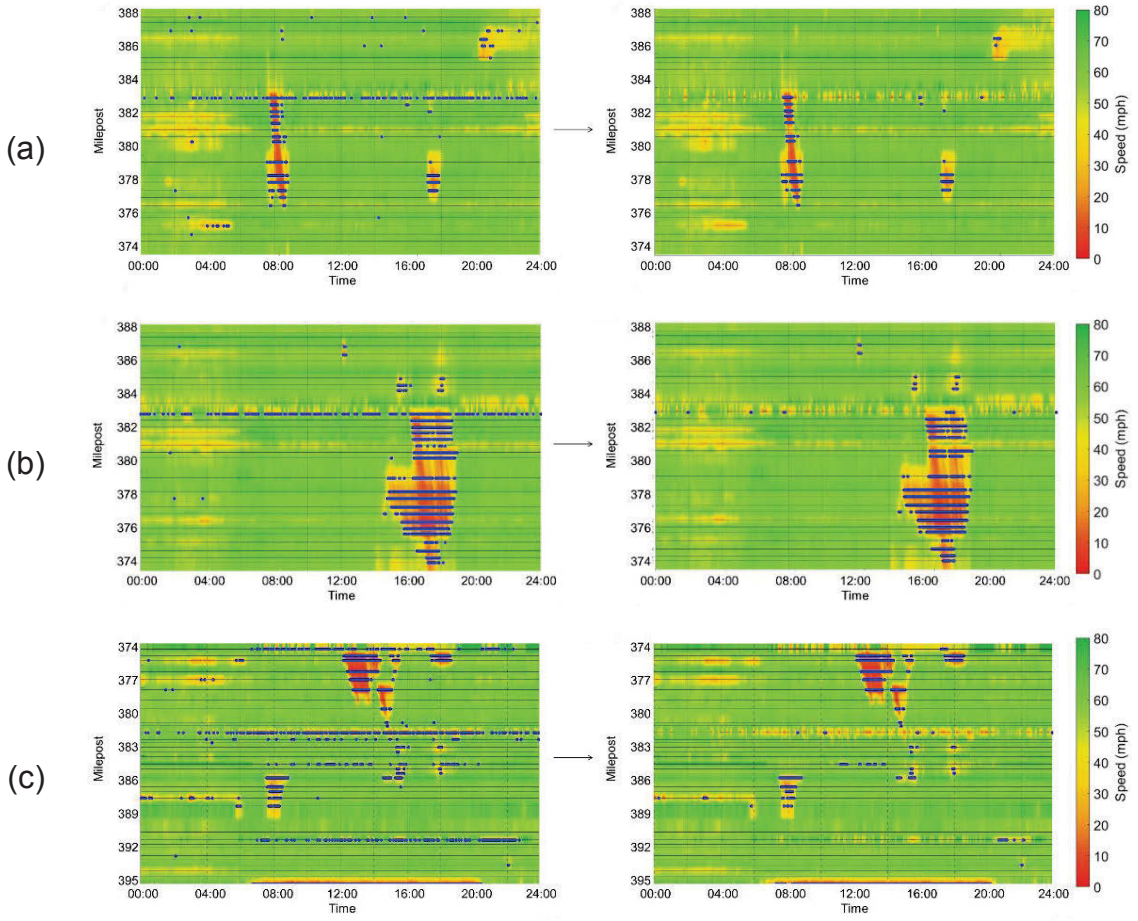


Figure 3-9 Phase identification and congestion detection results with speed heat map: (a) August 30, 2016, and (b) September 1, 2016 (I-40 EB), and (c) September 1, 2016 (I-40 WB).

detection rates over speeds. Figure 3-10 shows the congestion detection rates for the seven test days. Please note that the initial speed threshold of 15 mph was applied to estimate the probability distribution of the congested flow at each station. There are not enough samples of near-to-stop traffic for most of the stations in the RTMS data if the threshold of 5 mph was applied along with the definition of a queued state in HCM 2010 [68]. Thus, the relaxed condition of 15 mph was used in this study. The proposed algorithm detects 100% of 0-5mph conditions and 99.9% of 0-15mph conditions in the seven-day test data. The dashed lines in Figure 3-10 represent the average detection rates of all test days as a reference. The different characteristics of congestion patterns of each day can be observed by comparing a given day's detection rates to the average rates. For example, the congestion of August 30 is more severe than that of August 8 because the congestion detection rates in the mid-speed range, 15-40mph, on August 30 are lower than those on August 8.

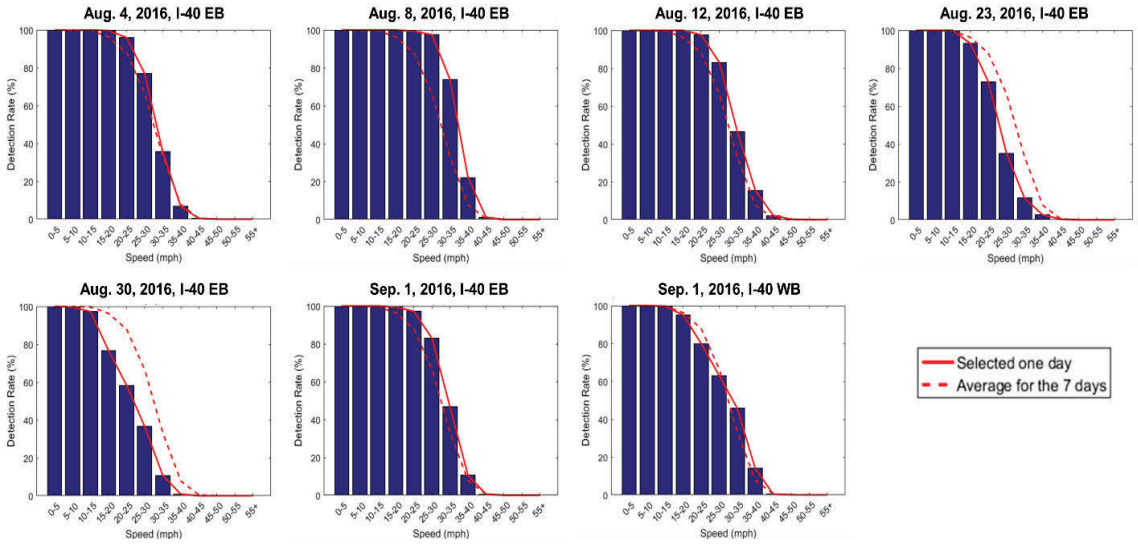


Figure 3-10 Congestion detection rate over speeds for each test day.

Shock wave speed calculation and queue prediction

The proposed algorithm aims to identify traffic flow phases by detecting the conditions of whether demand exceeds capacity. These conditions are threefold: (a) varying capacity over time with fixed demand, (b) varying demand over time with fixed capacity, and (c) mixed conditions of both (a) and (b). Typically, the first condition is the case of a nonrecurring incident such as a crash, where a backward recovery shock wave is observed (Figure 3-11(a)), while the second is peak-hour traffic conditions in which a forward recovery shock wave is formed (Figure 3-11(b)) [33].

In this study, the speeds of all the backward forming shock waves that extend across at least four upstream stations were tracked and calculated by using two empirical approaches, as explained in the previous section (Figure 3-5). A total 190 speeds were calculated from the 16 shock waves. Table 3-1 shows the comparison of the shock wave speeds from both approaches. For most cases, the shock wave speeds of the second approach were slower than those of the first approach because the second one tracks the “average” travel speed of the upstream front of a growing queue. In addition, the average shock wave speed of 11.6 mph from the second approach appears to be closer to the range of 15-20 kilometer per hour (kph) (equivalent to 9.3-12.4 mph) observed by other researchers [87, 89].

By assuming that the shock wave detected at the current station and time continues to travel to the next upstream station at the same speed, its arrival time at the next station was predicted. The prediction errors are measured using a mean absolute error as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|t_i - \hat{t}_i| \qquad (32)$$

where, t_i is the detected arrival time of a queue i and $\hat{t}_i$ is the predicted arrival time of the queue i.

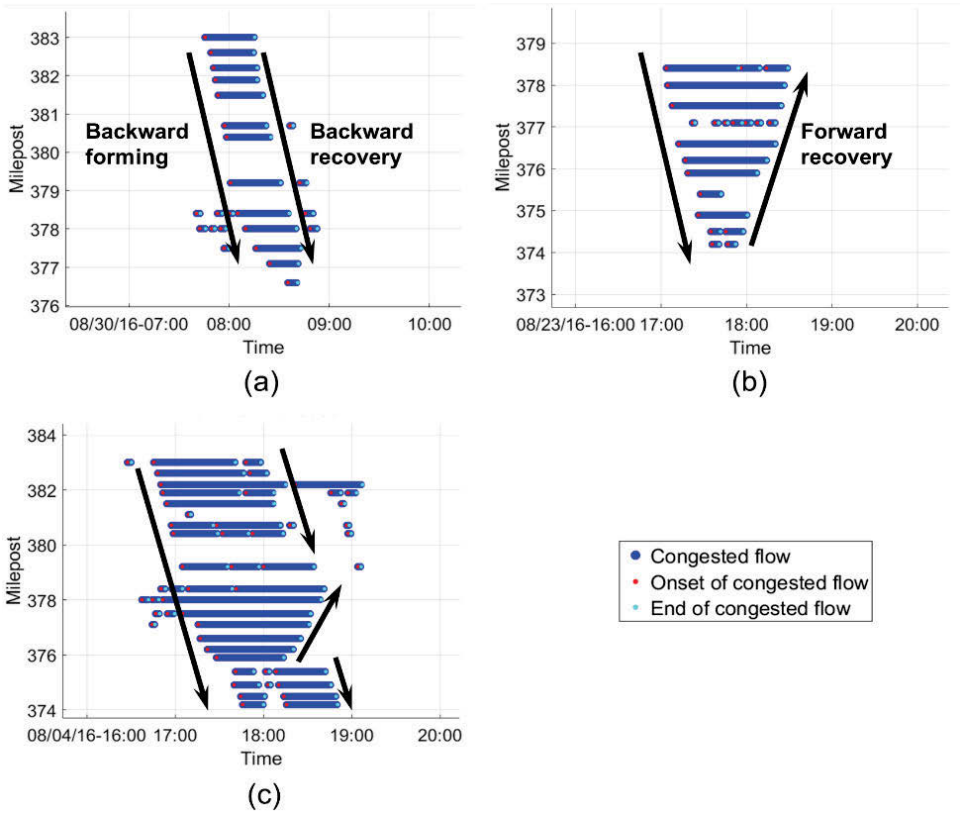


Figure 3-11 Shock wave examples based on congestion detection results: (a) varying capacity, (b) varying demand, and (c) mixed condition.

The prediction errors are shown in Table 3-1. On average, the MAE of the first approach is 162 seconds while that of the second approach is 149 seconds. In a comparison of the congestion types, the backward forming-backward recovery type has the smallest error, 133 seconds for both approaches. This implies that the shock wave speeds of a queue are relatively constant over time and location. For the other types, the second approach outperforms the first one. This is mainly because the shock wave speeds of a given queue fluctuate more over time and location compared to the first type. The backward forming-forward recovery type is often observed during peak hours due to varying demand conditions. Although the capacity at each station remains the same over time, the empirical shock wave speed fluctuates due to the capacity difference over locations. For such cases, the second approach can produce better predictions.

Conclusion

This paper proposes a real-time queue detection algorithm by using traffic flow fundamentals combined with a statistical pattern recognition procedure. First, the traffic phase identification procedure is applied to detect congested flows at each detector station where demand exceeds its capacity so that a queue is formed. GMMs of the traffic flows are estimated using historical flow density data to capture location-specific flow-density patterns. Then, new data are classified in a probabilistic manner into either a congested or uncongested flow phase, based on the estimated GMMs. Next, the congestion in a time-space domain is detected by collectively using the onsets and ends of the congested flow phase at each station. Finally, empirical shock wave speeds between two stations are calculated and the queue arrival time at the next upstream station is predicted.

This algorithm detected most of the low-speed conditions in the test datasets successfully, although it aims to detect the traffic conditions where

Table 3-1 Empirical shock wave speeds and queue arrival time prediction errors.

Category		Average Shock Wave Speed (mph)		MAE of Prediction Result (sec)	
		Approach 1	Approach 2	Approach 1	Approach 2
By days	Aug. 4	-9.7	-10.0	125	105
	Aug. 8	-12.0	-10.1	89	108
	Aug. 12	-14.0	-11.2	135	117
	Aug. 23	-21.4	-13.6	196	186
	Aug. 30	-14.1	-11.4	76	101
	Sep. 1 (EB)	-13.0	-11.6	222	188
	Sep. 1 (WB)	-19.0	-19.7	322	311
By types	Backward forming – backward recovery	-12.5	-11.2	133	133
	Backward forming – forward recovery	-19.1	-12.8	169	159
	Mixed	-12.2	-11.3	182	156
Total Average		-13.7	-11.6	162	149

demand exceeds capacity rather than identify low-speed conditions below a certain threshold. The different detection rate distributions with respect to speed range between the seven-day test datasets are the evidence that this algorithm is adaptive to varying queueing conditions and queue types over time and space. Two empirical approaches for calculating the shock wave speed between two stations were tested based on whether the downstream station of a queue is fixed. For a moving queue, typically caused by incidents and having the feature that its forming and recovery shock waves are backward, both approaches show very similar prediction performance. For a stationary queue, typically observed with a backward forming and forward recovery shock waves during peak hours, using shock wave speeds from the first downstream station predicts the queue arrival time at the next station better than using speeds between two neighboring stations.

Further research is needed to advance the sophistication of the prediction procedure in the proposed algorithm by accounting for additional variables, such as the flow or densify differentials between stations along a highway. It is more desirable to improve the prediction performance based upon traffic operational goals and strategies of traffic operations authorities. It is also useful to combine the proposed queue detection algorithm with an automatic incident detection algorithm so that the detected queue can be identified as either recurring or nonrecurring.

CHAPTER IV

GRAY AREAS IN ISOLATED INTERSECTION CONTROL-TYPE SELECTION: A COMPLEMENTARY DECISION-SUPPORT TOOL

A version of this chapter was originally published by Bumjoon Bae, Brandon C. Whetsel and Lee D. Han:
Bumjoon Bae, Brandon C. Whetsel, and Lee D. Han. “Gray Areas in Isolated Intersection Control Type Selection: A Complementary Decision-Support Tool.” *Journal of Transportation Engineering, Part A: Systems* 143(11) (2017): 04017055.

Abstract

The intersection control-type for future facilities can be determined by comparison of the common measure of effectiveness, average control delay. However, rigid comparisons of such measures tend to mislead the decision-making process in practice, since there must be latent factors in quantification. To this end, this paper proposes the performance comparison framework of different transportation facility alternatives using a common quantitative measure. By considering the uncertainties in a quantification process, the proposed framework provides gray areas, intuitively visualized information, that decision makers can use to assist their engineering judgement. The average control delay of two-way stop control, all-way stop control, signal control types, and roundabouts were compared with contour lines of delay differences. It is found that the delay of a roundabout increases rapidly as the traffic demand increases. Hence a signal control type has the minimum delay level in that case, despite the roundabout outperforms for most of the low-demand conditions. When the signal timing plan was optimized, this feature becomes remarkable. With consideration of the margin of error in the delay, a gray area on the minimum delay surface between the signal control and roundabout types enlarges in the low-demand area. The gray areas can be utilized by practitioners to decide the best

intersection control type with consideration of construction and maintenance costs over delay reduction benefit.

Introduction

Control delay is not only used to characterize the performance of each intersection control type but is also employed as a criterion for comparing the different types to each other. Besides the traffic signal warrants described in the *Manual for Uniform Traffic Control Devices* (MUTCD) [90], comparing the delay directly between different intersection control types is necessary. However, decision making based solely on quantitative metrics tends to be misguided toward quantitative fallacy. To address this issue in intersection control type selection, considerations must be taken for the factors which are not included in the control delay calculation process including: varying traffic demand, errors in model parameters and input data, user discernible delay margin, traffic growth and maintenance cost for the future, and so on.

Previous studies of two-way stop control (TWSC), all-way stop control (AWSC), signal control, and roundabouts have provided comparisons of efficiency and safety issues both qualitatively and quantitatively [91] [92]. Han, Li [91] employed the *Highway Capacity Manual* 2000 (HCM 2000) methodologies in order to compare the delay levels under TWSC, AWSC, and signal control types with varying demand and left-turn percentages. They proposed a minimum-delay surface with delineating curves, accounting for the control delay only, that distinguished the mutually exclusive minimum delay zones for the control types.

This chapter proposes the comparison framework of the average control delay under TWSC, AWSC, signal control, and roundabouts using the *Highway Capacity Manual* 2010 (HCM 2010) procedures. While accounting for the latent factors, a range of delineating curves is identified to distinguish each minimum

delay zone so that the “gray areas”, the overlapping areas defined by the range where the difference in the delay from each pair of the control types is marginal, can provide more flexibility to the related decision-making process.

The remainder of this paper is organized as follows. The next section provides background on potential errors in control delay estimation. Then, the intersection delay models of HCM 2010 [68] for TWSC, AWSC, signal control, and roundabouts are briefly reviewed with their features in the third section. The readers who are familiar with the *Highway Capacity Manual* (HCM) delay models may want to skip this section. The fourth section describes the design of case scenarios. The delay comparison results and the gray area between the control types are described and discussed in the fifth section. Finally, conclusions are drawn.

Background on Errors in Control Delay Estimation and Gray Areas

The results of the control delay calculation can be affected by uncertainty in model structure or required input data. Due to this uncertainty, relying on a rigid value of the minimum delay may lead to an inaccurate decision. For this section, the limitations of the HCM delay model and its significant factors affecting the resultant average control delay are reviewed from other studies.

The current control delay model for signalized intersections, in use since HCM 2000 [93]), is composed of the uniform, incremental, and initial queue delays. This model originated from Fambro and Rouphail [94]. They proposed a generalized delay model to account for actuated signal control parameters, oversaturation, variable demand, and metering and filtering effects from upstream traffic signals. For undersaturated conditions, the average delay mainly comes from the uniform delay which depends on a progression adjustment factor

(PF) by arrival type for a movement group. Benekohal and El-Zohairy [95] claimed that the HCM uniform delay model for coordinated signalized intersections is inaccurate due to the PF when either 1) a dense platoon arriving at the start of the green or red interval, or 2) a moderately dense or dispersed platoon arriving during the green interval. These will influence the average delay significantly.

Unlike the uniform delay, the non-uniform delay including the incremental and initial queue delays are determined by random arrivals and queues which contribute more under oversaturated conditions. Sazi Murat [96] argued the uncertainties of the variables in the HCM delay model especially for the non-uniform arrival or oversaturated condition despite the many efforts from other studies to alleviate randomness in the average delay. The author proposed the Neuro Fuzzy Delay Estimation model and Artificial Neural Networks Delay Estimation model. In a similar vein, Tian, Urbanik [97] showed that the highest variation in delay occurred when traffic demand approaches capacity and a range of speed variations has a high impact on the delay variation based on a simulation analysis.

The base saturation flow rate is another important factor to explain the control delay. Khatib and Kyte [98] argued that change in traffic volume or saturation headway has significant effects on delay variations. Taking into consideration the fact that the delay model is basically a function of demand and capacity, their finding is not surprising. Tarko and Tracz [99] claimed that the existing saturation flow prediction equation has a high standard error of 8-10% based on previous studies [100]. They also emphasized three sources of the errors in vehicle delay: temporal variance of a saturation flow; omitted capacity factors; and an inadequate functional relationship between model variables and saturation flow rates.

In general, there is a tradeoff between bias and variance in a quantification process where the closest measurement to its true value is

desired. The bias and variance can be substituted with accuracy and precision of the process, respectively. Increasing model precision can reduce its accuracy if there is a certain amount of uncertainty in the input values [101]. Han, Li [91] proposed rigid delineating curves composed of traffic volumes on major and minor streets to identify the best intersection control type which has the minimum control delay. Those results may be precise but not accurate since traffic demand fluctuates and other traffic conditions are varying spatiotemporally so that the delay from the models may have errors. More accurate results can be obtained by loosening the precision, which is applying a range of delineating curves, instead of rigid lines.

In this study, the range of the delineating curves, i.e., gray area is identified to address the error in control delay, attributed to all the influential factors above for TWSC, AWSC, signal control, and roundabout types under given traffic conditions. The way of comparison using gray areas is helpful for practitioners to make a more accurate decision. However, there have not been enough studies in a literature to identify the amount of errors in the resultant control delays. 8-10% error in the base saturation flow rate based on Tarko and Tracz [99] results in variability of the control delay. For example, the resultant control delay of signal control type varies ranging from -2.4 seconds (-9%) to 4.4 seconds (16%) when $\pm$10% change of the saturation flow rate under the given conditions in this paper. Thus, the gray area is assumed as a $\pm$5-second difference in the control delay between the best and second-best control types in this study. Note that it is recommended for practitioners to set their own gray area ranges depending on the purpose and type of the decision makings.

HCM Intersection Delay Models

According to HCM 2010 [68], control delay is the resultant delay caused when a traffic movement reduces speed or stops due to a traffic control device. Therefore, it represents the additional travel time over the uncontrolled condition [68], [93]. This definition is consistent in signalized and unsignalized intersections as well as roundabouts [91], such that this measure can be used for comparison of the performance and level of service (LOS) between the three control types. The delay models for traffic signal and stop sign control types in HCM 2010 [68] are identical, for isolated intersections, with those in HCM 2000 [93]. The delay model for roundabouts had been newly added in the 2010 edition.

Signalized Intersections

The average control delay d for signalized intersections is composed of three sub components, expressed by Equation (33).

$$d = d_1 + d_2 + d_3 \tag{33}$$

Where, d_1 is uniform delay occurring when uniform arrivals are assumed, d_2 is incremental delay including delay due to random arrivals and cycle failures during the analysis time period, and d_3 is initial queue delay experiencing all vehicles in the analysis period due to an initial queue presenting at the start of the analysis period.

Since an initial queue was assumed not to exist for this study, the average control delay, d, can be expressed by Equation (34).

$$d = \frac{0.5C(1-g/C)^2}{1-[\min(1,X)g/C]} + 900T\left[(X-1) + \sqrt{(X-1)^2 + \frac{8kIX}{c_{lg}T}}\right] \tag{34}$$

where, C is the cycle length (s), g equals effective green time (s), X is the volume-to-capacity ratio or degree of saturation, T is the analysis period duration

(h), k is the incremental delay factor, I equals upstream filtering adjustment factor, and c_{lg} is the lane group capacity (veh/h).

Two-Way Stop Control (TWSC)

The procedure to measure average control delay for TWSC, in HCM 2010, is based on field measurements in the U.S. with a gap acceptance model that was developed in Germany [68]. Since only minor street approaches are controlled by stop signs in TWSC intersections, the control delay is not defined for the major street. According to HCM 2010 [68], the average control delay for any minor movement of TWSC intersections is expressed by Equation (35). As in the case of the signal control type, the average control delay is the function of the capacity and the degree of saturation. Both factors can be affected significantly by the conflicting flow rate for each movement on the minor street due to the fact that this procedure relies on the gap acceptance model. Therefore, as the traffic volume on the major street approaches capacity, unrealistically large values of delay can be observed.

The constant term, 5 s/veh explains the time to slow down from free-flow speed, stop, then accelerate to free-flow speed for a vehicle on the minor street.

$$d = \frac{3{,}600}{c_{m,x}} + 900T\left[\frac{v_x}{c_{m,x}} - 1 + \sqrt{\left(\frac{v_x}{c_{m,x}} - 1\right)^2 + \frac{\left(\frac{3{,}600}{c_{m,x}}\right)\left(\frac{v_x}{c_{m,x}}\right)}{450T}}\right] + 5 \qquad (35)$$

where, $c_{m.x}$ is the capacity of movement x (veh/h), v_x equals the flow rate for movement x (veh/h), and other variables are the same in the previous equations.

All-Way Stop Control (AWSC)

The average control delay for AWSC in HCM 2010 [68] is calculated by an iterative procedure with three key time-based terms: the saturation headway, the departure headway, and the service time [68]. The headways rely on the degree

of conflict between consecutively departing vehicles on the subject approach and the vehicles on other approaches. The number of vehicles conflicted by the subject vehicle and the number of lanes on the intersection approaches are the main factors of the degree of conflict. Because capacity for AWSC is equal to the maximum throughput on an approach, under the given traffic flow rates on the other approaches, it can be concluded that the traffic demand and base headway assumptions given in the procedure are important components for the control delay. The average control delay for AWSC in HCM 2010 [68] is expressed by Equation (36).

$$d = t_s + 900T\left[(x-1) + \sqrt{(x-1)^2 + \frac{h_d x}{450T}}\right] + 5 \quad (36)$$

where, t_s equals the service time (s), which is average time spend by a vehicle in first position waiting to depart, h_d is the departure headway (s).

Roundabout

The capacity of a roundabout approach heavily relies on the conflicting flow rate, which represents the circulating flow faced by the subject approach vehicles. The functional form of the average control delay model for a roundabout in HCM 2010 [68], expressed by Equation (37), is identical with that for TWSC except for the 5-second constant term. The additional delay assumption is loosened for a roundabout accounting for the YIELD control on the subject approach in undersaturated conditions.

$$d = \frac{3{,}600}{\mathrm{c}} + 900T\left[(x_{vc}-1) + \sqrt{(x_{vc}-1)^2 + \frac{\left(\frac{3{,}600}{\mathrm{c}}\right)x_{vc}}{450T}}\right] + 5 \times \min[x_{vc}, 1] \quad (37)$$

where, c is the capacity of the subject lane (veh/h), x_{vc} equals the volume-to-capacity of the subject lane.

From the aforementioned average control delay models for signalized, stop controlled intersections, and roundabouts, it can be concluded that the major factors affecting the level of control delay are traffic volume, i.e., demand and capacity. Therefore, this study accounts for both factors in a sensitivity analysis framework.

The capacity of each control type is however based on assumptions in which base saturation flow rates (or base saturation headways) are different between each type. For example, as a default value, the base saturation flow rates (pc/h/ln) are 1,900 for each movement (Signal Control), 1,700 for movements on the major street (TWSC), 923 for the degree-of-conflict case 1 (ASWC), and 1,130 (Roundabouts) [68]. Therefore, the difference in values of the control delay of every pair among the four control types is explored directly in this study, assuming it is mainly caused by the varying capacity. For the demand side, the different major and minor street traffic volumes as well as different percentages of left-turn traffic volumes were considered in the comparison of the control delay. More details for the scenario design of this study are described in the following section.

Design of Case Scenarios

In the context of the objective of this study, 4,305 cases for each intersection control type plus signal timing optimization scenarios, totaling 21,525 cases, are analyzed in terms of the major and minor-street volumes as well as the percentage of left-turn traffic volumes using HCM 2010 [68]. The geometric design, traffic, and signal parameters are applied for the simplest and generic manner. This is consistent with the previous study, Han, Li [91], so that the

results can be compared to each other. The same analysis time duration of 15 minutes is used for all cases.

Intersection Configuration

A simple isolated 4-legged intersection where each approach has a single lane is used in this study. Each lane on the approaches is assumed as a shared left turn, right turn, and through lane. All related parameters are accounted for as default values from HCM.

Traffic Demand

Traffic demand for the major and minor streets range from 0 to 2,000 veh/h in 50 veh/h increments. Each increment is analyzed using 5 different levels of left turns: 0%, 5%, 10%, 15%, and 20% of total demand on each approach. The cases where the minor street demand exceeds the major street demand are excluded in the analysis. Therefore, a total of 4,305 traffic demand cases are applied for each control type. No bike or pedestrian demand is considered in this study. The cases where a volume on the minor street exceeds that on the major street are excluded in order to keep the hierarchy for both streets.

Traffic Control Parameters

All the traffic control parameters (e.g., base critical gap, saturation headway, etc.) included in TWSC and AWSC are applied as the default values from HCM. The cycle length is 60 seconds and the phase splits are assigned as 50/50. The yellow change and red clearance time is set as 4 seconds. For the signal optimization scenarios, the cycle length is computed by Equation (38) in HCM 2010 [68].

$$C = \frac{L}{1 - \left(\frac{CS}{RS}\right)} \tag{38}$$

with

$$C_{min} \leq C \leq C_{max}$$

and

$$C = C_{max} \text{ when } CS \geq RS$$

where L is the cycle lost time (s), $L = 8$, CS is the critical sum (veh/h), RS equals the reference sum flow rate (veh/h), $RS = 1{,}530 \times PHF \times f_a$, PHF is the peak hour factor, f_a is the adjustment factor, C_{min} and C_{max} are the minimum and maximum cycle length.

After the cycle length is determined in each scenario, the splits are optimized to minimize the average delay of the intersection.

Analysis Results and Comparisons

Results of the average control delay were obtained from the 4,305 cases of each control type. Figure 4-1 demonstrates the contours of average control delay for each control type at 20% left-turn volumes. One can identify which type has the minimum delay under a given demand level from the contours. It is obvious that the delay levels of a roundabout are lower than the other types, particularly for lower volumes on both major and minor streets. The spacing between the contours of each control type and the direction of contours show how fast the control delay increases as traffic volumes on either the major street or minor street increase. For example, the 10-50 second delay contours of TWSC are densely plotted and move along the vertical axis. This indicates that the control delay of TWSC increases rapidly as the minor street volume increases. In contrast, the signal control delay shows a relatively slow increase as the major street volume increases. The delay patterns of the roundabout are similar with that of the signal control. However, the roundabout is more sensitive to the minor street volume than signal control.

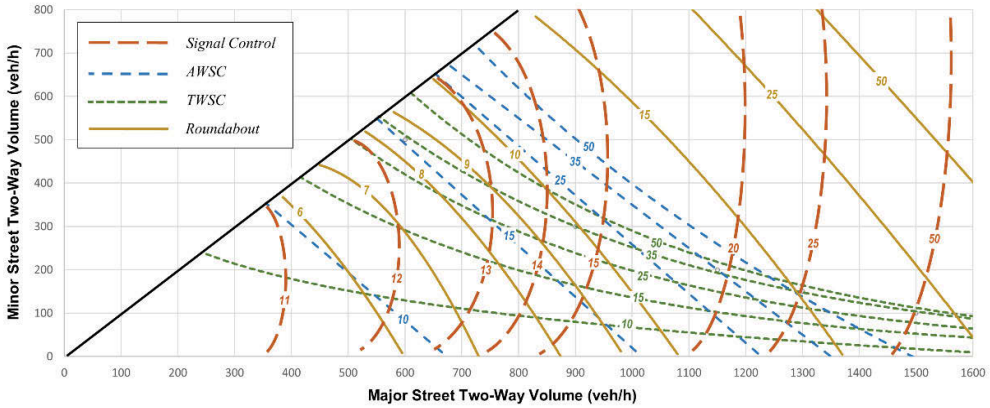


Figure 4-1 Contours of control delay for signal control, AWSC, TWSC, and roundabout with 20% left turns.

A 3-D surface plot can show how the delay of each control type changes depending on either the major or minor street volume and make the delay comparison much clearer. Figure 4-2 shows the control delay surfaces with 20% left-turn volumes for the intersection control types. Because the vertical axis representing average control delay is flipped over, the surface on top represents the control type having the minimum control delay with given traffic conditions. As mentioned above, a roundabout performs best for most of the demand area. The surface of signal control emerges above that of a roundabout as the volume on the major and minor streets increase. TWSC performs best when the minor street volume is very low. However, its performance heavily relies on the delays on the minor street. Thus, the average control delay of TWSC rapidly increases as the minor street volume increases. Although AWSC does not show up for any demand levels in Figure 4-2, it is good to know where the delay surface of AWSC is located and how it looks under the other surfaces in order to understand its gray areas. As illustrated in Figure 4-3, in the comparison of TWSC, AWSC, and signal control without a roundabout, AWSC performs best when traffic volumes on the major and minor streets are somewhat balanced, and the total volume is less than 900 veh/h.

Figure 4-4 shows the minimum delay surfaces with gray areas represented by contour lines indicating the $\pm$5-second difference in the delay between the best and second-best control types. The black solid lines are the delineating curves where the delays of both control types are equal. In order to understand how the gray areas enlarge or shrink at each LOS, the LOS regions are also depicted for each control type.

For the demand area over 800 veh/h on the major street and over 500 veh/h on the minor street, signal control type is the best in terms of the average control delay showing LOS B through E for the 20% left-turn scenario. A roundabout shows the best performance for the lower volume area. However, the minimum traffic volumes on both single-lane major and minor

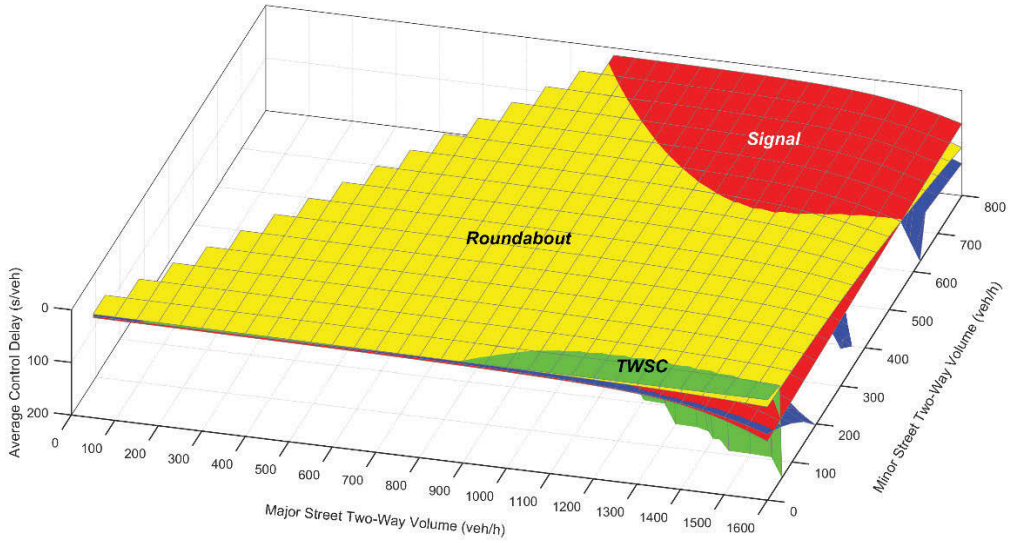


Figure 4-2 Delay surfaces for 4 different control types, with 20% left turns.

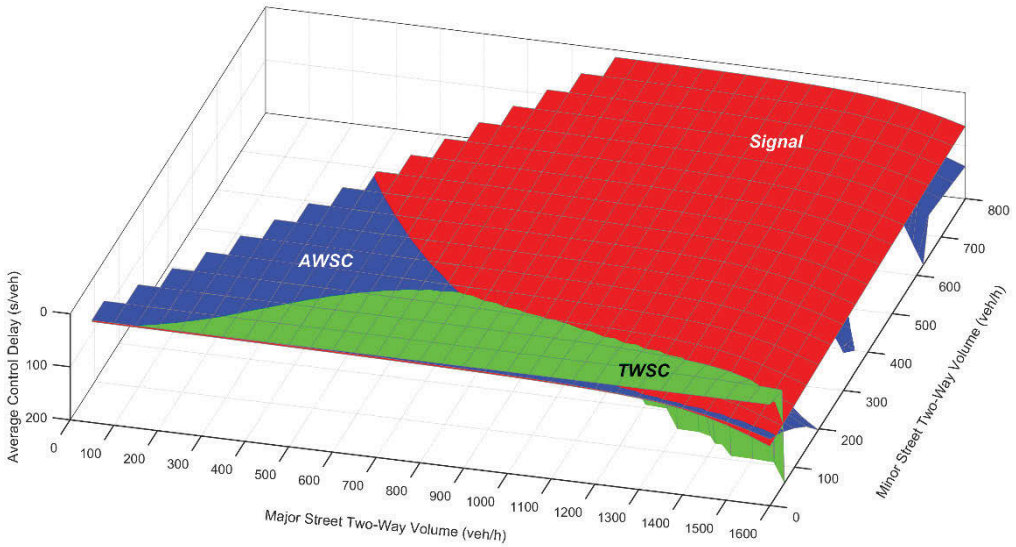
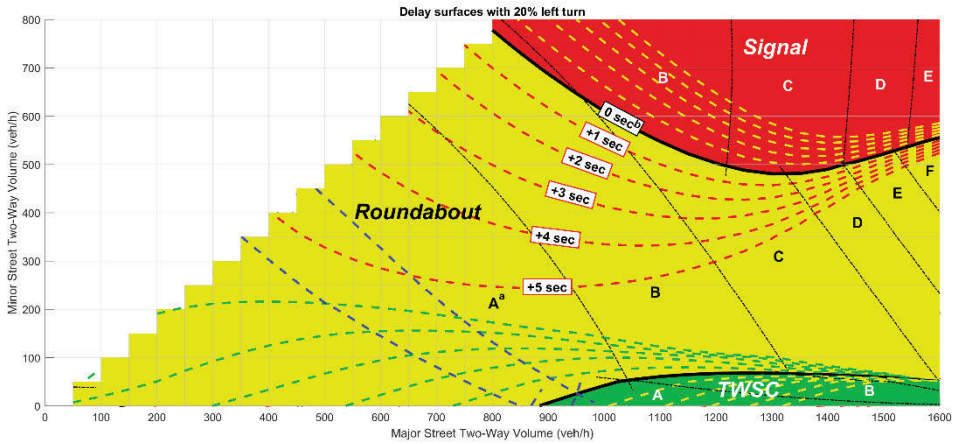


Figure 4-3 Delay surfaces for 3 different control types, with 20% left turns.



[a] Upper case letter indicates the level of service for each control type;
[b] Each solid and dotted contour line represents 0-5 second difference in delay between two control types.

Figure 4-4 Comparison of delay by control types and gray zones with 20% left turns.

streets of the traffic signal warrant 1 in MUTCD [90] are 500 veh/h and 150 veh/h, respectively. The large gap of the traffic volume thresholds between the warrant 1 and the result in Figure 4-4 implies that huge room for engineering judgement exists to select either signal control or a roundabout. If the gray area between two control types was taken into account, better judgement can be made based on the additional information of how different the delays are under a given traffic condition. When an observed traffic condition falls onto the gray area, a practitioner may want to choose the second best control type in terms of delay. For example, under 800 veh/h on the major street and 300 veh/h on the minor street, the best performing control type can be signal control, not a roundabout. In contrast, a roundabout may still outperform signal control under the condition of 1,000 veh/h and 700 veh/h on the major and minor street, respectively. In the same way, TWSC can be the optimal control type for 400 veh/h and 200 veh/h on the major and minor street respectively instead of a roundabout. Although the AWSC area did not show up on the figure, 4- and 5-second delay difference contour lines against a roundabout appeared when the total traffic volume is less than around 1,000 veh/h.

Another noticeable feature is that the size of the gray areas on both sides of the solid line is asymmetric. For example, in a comparison between a roundabout and signal control, the size of the gray area on the roundabout is much larger than that of the signal control. Similarly, the gray area between a roundabout and TWSC spreads out more on the roundabout surface. This implies that even if a roundabout outperforms the others for a certain traffic volume range, its efficiency is marginal especially for LOS A through LOS C conditions.

Figure 4-5 displays the delay comparison results with no left-turn, 5%, 10%, and 15% of left-turn volume scenarios. As the percentage of left-turn volume increases, surfaces moves to the left as a whole. In addition, the delay of all control types increases so that the LOS range extends from A-C to A-E (A-F in

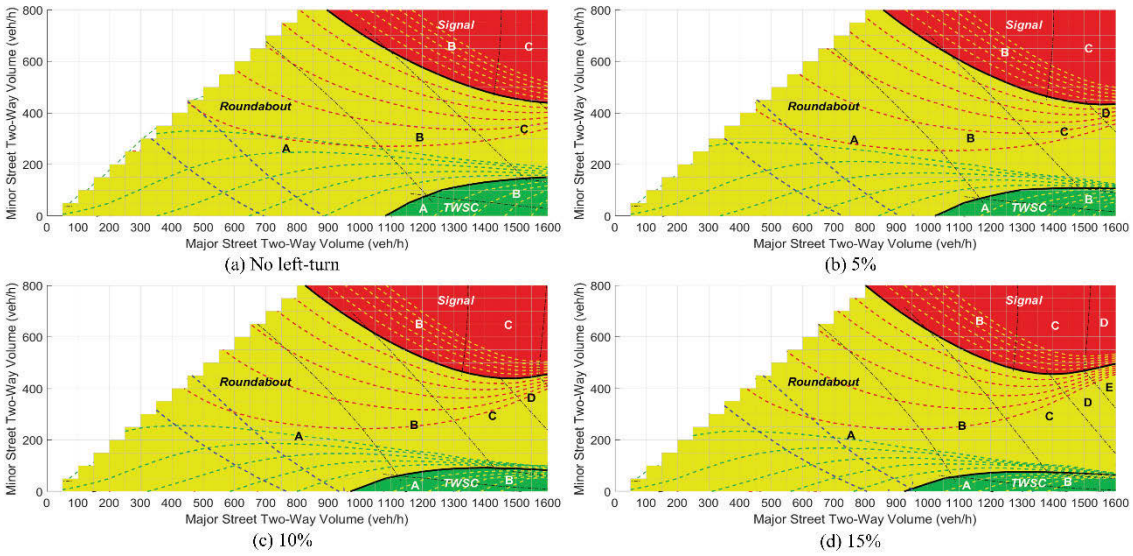


Figure 4-5 Comparison of delay by control types and gray zones with 0%, 5%, 10%, 15% left turns.

Figure 4-4). When the left-turn volume increases, the signal control area moves to the left slowly and the TWSC area shrinks down slightly. The corresponding gray area of each control type also moves along the delay surface.

For the signal control type, the signal timing plan was optimized depending on traffic volumes to minimize the average delay. Figure 4-6 and Figure 4-7 illustrate the delay comparison with 0-20% of left-turn volume including the optimized signal control type. In comparison with the non-optimization scenarios above, the delay of signal control type was substantially reduced so that its surface emerged for the range of over 950 veh/h on the major street indicating LOS A and B in the 20% left-turn scenario. In addition, optimized signal control outperforms TWSC as well, the TWSC surface does not appear anymore. However, the gray area on the signal surface is extended up to around 1,200 veh/h on the major street indicating the delay difference with a roundabout is still marginal. Similarly, most of the demand area less than the major street volume of 950 veh/h, where a roundabout mostly has the minimum control delay, is covered by gray areas. In conclusion, most of the demand area of approximately 1,200 veh/h or less is in the gray areas of all four control types. As the percentage of left-turn volume increases, the delineating curve and gray areas shift to the left. That is, the performance of a roundabout diminishes gradually because the control delay of a roundabout is significantly affected by conflicts between the circulating traffic and approaching traffic.

Conclusion

This study proposed the performance comparison framework of different transportation facility alternatives using a common quantitative measure. By considering uncertainties in a quantification process, the proposed framework provides the gray areas in such complementary comparisons so that it assists in

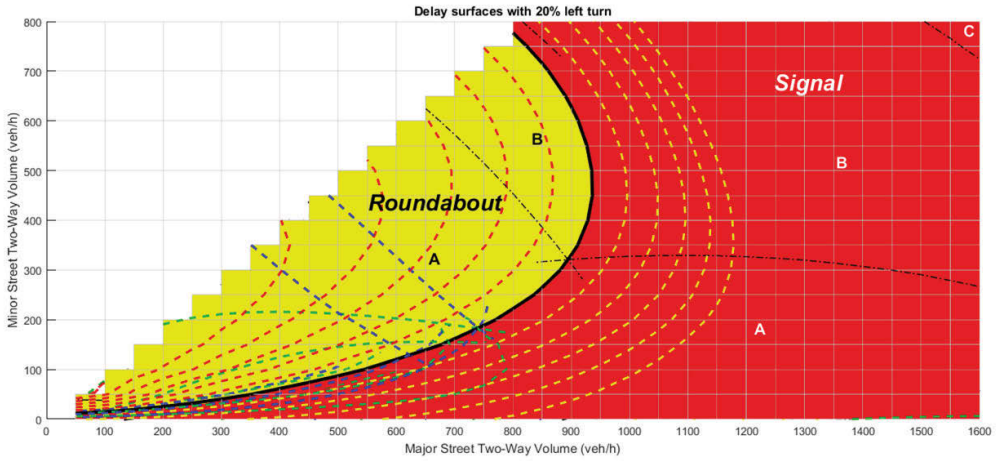


Figure 4-6 Comparison of delay by control types with signal optimization and gray zones with 20% left turn.

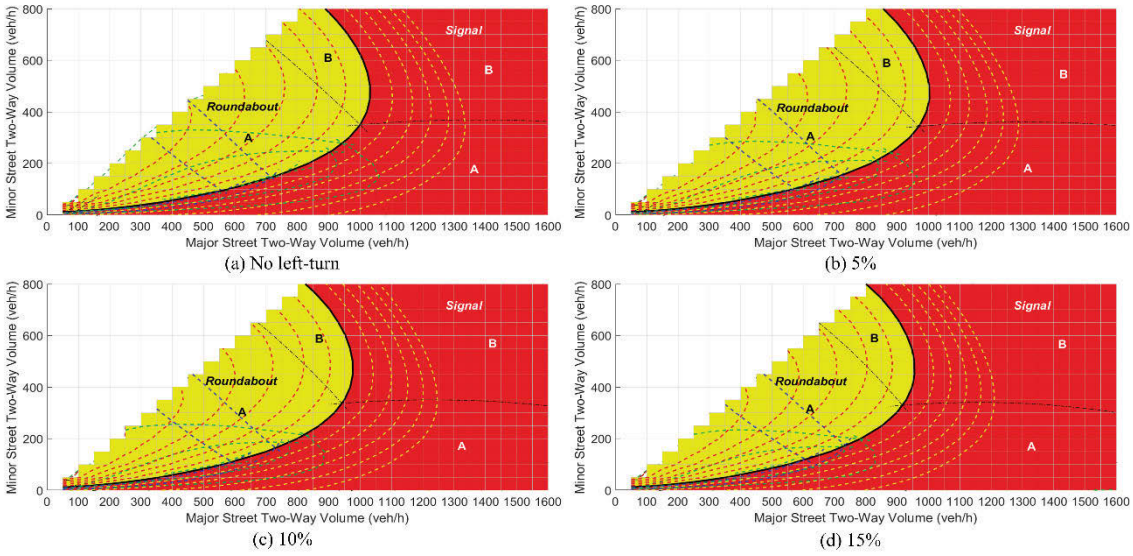


Figure 4-7 Comparison of delay by control types with signal optimization and gray zones with 0%, 5%, 10%, 15% left turns.

making a decision for transportation facility type selection.

For an isolated intersection control type selection, the HCM 2010 procedures to measure average control delay were employed and the performances of TWSC, AWSC, signal control, and roundabouts under given traffic conditions were compared.

Conservatively it can be concluded that a roundabout outperforms the others for most of the cases when the minor street two-way traffic volume is less than 400 veh/h and the left-turn volume percentage on each approach is less than 20%. However, the control delay of a roundabout increases more rapidly beyond these demand ranges such that signalized control emerges on top in the minimum control delay surface plot.

Both stop-control types have a higher control delay level for most cases, although the differences in the minimum delay for a roundabout are marginal such that the performances of TWSC and AWSC are not significantly worse for the relatively lower demand conditions. TWSC shows the best performance with a high major street volume and very low minor street volume. Caution is required to interpret this feature. It is not only indicating the minimum overall intersection delay but also implies severe delay on the minor street because the measure used for these comparisons is an “average” control delay which is weighted by the traffic volume on each approach.

The control delay derived from the HCM 2010 approaches relies heavily on traffic volume and corresponding capacity levels. That is, the resultant delay can be different due to potential errors in input volumes and/or model parameter values. To this end, the study identified a gray area defined by the delay-difference contour lines between two control types for the purpose of a sensitivity analysis. As expected the gray area is larger for the lower demand conditions, which implies that it can provide practitioners more room for so-called “engineering judgement”.

Selecting and installing a roundabout entails a trade-off. Since its geometric design characteristics are considerably different than a conventional intersection design, once installed, it would be cost prohibitive to convert it to another control type. Therefore, it is recommended to assign more priority to signal or stop control in cases where a rapid growth in traffic demand is expected in the near future and the total delay-reduction benefit is expected to be lower than the potential cost in the future.

Signal optimization requires the use of actuated signal operations that respond to fluctuations in traffic demand. Such operating systems come with high installation and maintenance costs for a small intersection with light traffic demand. Practitioners should carefully consider the cost-effectiveness in utilizing signal optimization.

For the purpose of forecasting a likely intersection control type for future facilities, this study focused on a simple isolated intersection with mostly undersaturated traffic demand and presumed “default” conditions provided by HCM 2010 [68]. Therefore it has the following limitations:

- Intersection configuration: Only a single-shared-lane-approach intersection was analyzed in this study. For the signal type, a permitted left-turn signal plan was used. Estimating capacity and control delay of a shared-lane and/or permitted left-turn case requires additional complicated procedures in HCM 2010 [68], hence the delay comparison results may be different for different intersection configurations.
- Bike and pedestrian demand: No consideration was assigned for the bike and pedestrian demands. The pedestrian volume can particularly affect the control delay of left-turn traffic movements.
- Traffic demand balance: The traffic volume on one and opposing approaches was assigned as 50/50 percentage for simplicity. For TWSC, AWSC, and roundabouts the delay on a subject approach is greatly affected by the traffic volume of the conflicting traffic movements from other

approaches. Therefore, the resultant control delay levels can be different for such unbalanced traffic demand conditions.

CONCLUSION

This dissertation compiled a series of studies on short-term traffic flow dynamics in a spatio-temporal domain and on uncertainty in a decision-making process to support real-time traffic operations. These studies were conducted to propose multiple applications to impute missing traffic data with a secondary data source, predict traffic speed for a large road network, detect traffic queues in real-time, and support an engineering judgement in evaluating the performance of traffic facilities.

First, three kriging-based spatio-temporal missing data imputation approaches were proposed with and without using a secondary data source and the performance was evaluated under different patterns of missing data. A simple cokriging method improved the accuracy of imputation when the missing pattern was not random. In contrast, using only primary data with ordinary kriging outperforms the cokriging methods when the missing pattern is completely random.

Second, a nonparametric data-adaptive traffic speed prediction algorithm was proposed. The algorithm effectively reduces the dimensionality of traffic speed data in a spatio-temporal domain and predicts the future speed accurately. The proposed algorithm outperformed the benchmark models in terms of prediction accuracy for abnormal traffic conditions with much shorter computation time.

Third, a real-time queue detection algorithm was developed based upon traffic flow fundamentals combined with a statistical pattern recognition method. The proposed algorithm accounts for varying capacity-demand conditions in spatio-temporal dimension and collectively detects a queue along a highway. Further study is recommended to advance the sophistication of the queue prediction function in the algorithm.

Finally, the concept of gray areas was proposed for making an engineering judgement in transportation planning, management, or operations. A case study on intersection control type selection was performed to identify and visualize the gray areas. The proposed concept and the result of the case study can give additional intuitive and visualized information of uncertainties in a quantification process for comparing the performance of multiple alternatives.

Altogether, this dissertation provides a real-time traffic analysis framework that consists of multiple algorithms and tools for traffic operations of highway facilities. In addition to improving these algorithms, additional studies on the development of an automatic incident detection algorithm and an online traffic simulation tool will give greater sophistication to the analysis framework.

REFERENCES

1. Buuren, S.v., *Flexible Imputation of Missing Data*. Interdisciplinary Statistics Series. 2012, Boca Raton London New York: CRC Press.
2. Carpenter, J.R. and M.G. Kenward, *Multiple Imputation and its Application*. 2013.
3. Orchard, T. and M.A. Woodbury. *A missing information principle: Theory and application*. in *Proceddings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*. 1972.
4. Smith, B.L., W.T. Scherer, and J.H. Conklin, *Exploring Imputation Techniques for Missing Data in Transportation Management Systems.* Transportation Research Record: Journal of the Transportation Research Board, 2003. **1836**.
5. Ni, D. and J. Leonard Ii, *Markov Chain Monte Carlo Multiple Imputation Using Bayesian Networks for Incomplete Intelligent Transportation Systems Data.* Transportation Research Record: Journal of the Transportation Research Board, 2005. **1935**: p. 57-67.
6. Bennett, R.J., R.P. Haining, and D.A. Griffith, *Review Article: The Problem of Missing Data on Spatial Surfaces.* Annals of the Association of American Geographers, 1984. **74**(1): p. 138-156.
7. Rubin, D.B., *Inference and missing data.* Biometrika, 1976. **63**(3): p. 581-592.
8. Qu, L., et al., *PPCA-Based Missing Data Imputation for Traffic Flow Volume: A Systematical Approach.* IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, 2009. **10**(3): p. 512-522.
9. Li, L., Y. Li, and Z. Li, *Efficient missing data imputing for traffic flow by considering temporal and spatial dependence.* Transportation Research Part C: Emerging Technologies, 2013. **34**: p. 108-120.
10. Tang, J., et al., *A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data*

estimation. Transportation Research Part C: Emerging Technologies, 2015. **51**: p. 29-40.

11. Duan, Y., et al., *An efficient realization of deep learning for traffic data imputation.* Transportation Research Part C: Emerging Technologies, 2016. **72**: p. 168-181.
12. Eom, J., et al., *Improving the Prediction of Annual Average Daily Traffic for Nonfreeway Facilities by Applying a Spatial Statistical Method.* Transportation Research Record, 2006. **1968**: p. 20-29.
13. Wang, X. and K. Kockelman, *Forecasting Network Data.* Transportation Research Record: Journal of the Transportation Research Board, 2009. **2105**: p. 100-108.
14. Zou, H., et al., *An improved distance metric for the interpolation of link-based traffic data using kriging: a case study of a large-scale urban road network.* International Journal of Geographical Information Science, 2012. **26**(4): p. 667-689.
15. Selby, B. and K.M. Kockelman, *Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression.* Journal of Transport Geography, 2013. **29**: p. 24-32.
16. Shamo, B., E. Asa, and J. Membah, *Linear Spatial Interpolation and Analysis of Annual Average Daily Traffic Data.* Journal of Computing in Civil Engineering, 2015. **29**(1).
17. Yang, J., et al., *A Spatio-Temporal Approach for High Resolution Traffic Flow Imputation*, in *Transportation Research Board 96th Annual Meeting*. 2016: Washington, D.C.
18. Marcotte, D., *Cokriging with matlab.* Computers & Geosciences, 1991. **17**(9): p. 1265-1280.
19. Al-Deek, H.M. and C.V.S.R. Chandra, *New algorithms for filtering and imputation of real-time and archived dual-loop detector data in I-4 data*

warehouse. Transportation Research Record: Journal of the Transportation Research Board, 2004. **1867**: p. 116-126.

20. Zhong, M., S. Sharma, and P. Lingras, *Genetically Designed Models for Accurate Imputation of Missing Traffic Counts.* Transportation Research Record: Journal of the Transportation Research Board, 2004. **1879**: p. 71-79.
21. Tan, H., et al., *A tensor-based method for missing traffic data completion.* Transportation Research Part C: Emerging Technologies, 2013. **28**: p. 15-27.
22. Song, Y. and H.J. Miller, *Exploring traffic flow databases using space-time plots and data cubes.* Transportation, 2011. **39**(2): p. 215-234.
23. Yang, J., et al., *Short-Term Freeway Speed Profiling Based on Longitudinal Spatiotemporal Dynamics.* Transportation Research Record: Journal of the Transportation Research Board, 2014. **2467**: p. 62-72.
24. Krige, D.G., *A statistical approach to some basic mine valuation problems on the Witwatersrand.* Journal of the Southern African Institute of Mining and Metallurgy, 1951. **52**(6): p. 119-139.
25. Srinivasan, B.V., R. Duraiswami, and R. Murtugudde, *Efficient kriging for real-time spatio-temporal interpolation*, in *20th Conference on Probability and Statistics in the Atmospheric Sciences*. 2010: Atlanta, Georgia.
26. Gräler, B., et al., *Spatio-temporal analysis and interpolation of PM10 measurements in Europe for 2009*. 2013.
27. Wackernagel, H., *Multivariate Geostatistics: An Introduction with Applications*. Third Edition ed. 2003, Verlag Berlin Heidelberg New York: Springer.
28. Cressie, N., *Statistics for Spatial Data*. 1993, New York: Wiley.
29. Yates, S.R. and A.W. Warrick, *Estimating Soil Water Content Using Cokriging1.* Soil Science Society of America Journal, 1987. **51**(1): p. 23-30.

30. Basaran, M., et al., *Spatial information of soil hydraulic conductivity and performance of cokriging over kriging in a semi-arid basin scale.* Environmental Earth Sciences, 2010. **63**(4): p. 827-838.
31. Yang, J., *Spatio-Temporal dynamics of short-term traffic*, in *Civil and Environmental Engineering*. 2015, University of Tennessee, Knoxville.
32. Olea, R.A., *Geostatistics for Engineers and Earth Scientists*. 1999, New York: Springer Science+Business Media.
33. May, A.D., *Traffic Flow Fundamentals*. 1990, New Jersey: Prentice Hall.
34. Alecsandru, C. and S. Ishak, *Hybrid Model-Based and Memory-Based Traffic Prediction System.* Transportation Research Record: Journal of the Transportation Research Board, 2004. **1879**: p. 59-70.
35. Ishak, S. and C. Alecsandru, *Optimizing Traffic Prediction Performance of Neural Networks under Various Topological, Input, and Traffic Condition Settings.* Journal of Transportation Engineering, 2004. **130**(4): p. 452-465.
36. Vanajakshi, L. and L.R. Rilett. *A comparison of the performance of artificial neural networks and support vector machines for the prediction of traffic speed*. in *IEEE Intelligent Vehicles Symposium, 2004*. 2004.
37. Yang, F., et al., *Online Recursive Algorithm for Short-Term Traffic Prediction.* Transportation Research Record: Journal of the Transportation Research Board, 2004. **1879**: p. 1-8.
38. Chandra, S. and H. Al-Deek, *Cross-Correlation Analysis and Multivariate Prediction of Spatial Time Series of Freeway Traffic Speeds.* Transportation Research Record: Journal of the Transportation Research Board, 2008. **2061**: p. 64-76.
39. Chandra, S.R. and H. Al-Deek, *Predictions of Freeway Traffic Speeds and Volumes Using Vector Autoregressive Models.* Journal of Intelligent Transportation Systems, 2009. **13**(2): p. 53-72.
40. Guo, J. and B. Williams, *Real-Time Short-Term Traffic Speed Level Forecasting and Uncertainty Quantification Using Layered Kalman Filters.*

Transportation Research Record: Journal of the Transportation Research Board, 2010. **2175**: p. 28-37.

41. Min, W. and L. Wynter, *Real-time road traffic prediction with spatio-temporal correlations.* Transportation Research Part C: Emerging Technologies, 2011. **19**(4): p. 606-616.
42. Ye, Q., W.Y. Szeto, and S.C. Wong, *Short-Term Traffic Speed Forecasting Based on Data Recorded at Irregular Intervals.* IEEE Transactions on Intelligent Transportation Systems, 2012. **13**(4): p. 1727-1737.
43. Vlahogianni, E.I., M.G. Karlaftis, and J.C. Golias, *Short-term traffic forecasting: Where we are and where we're going.* Transportation Research Part C: Emerging Technologies, 2014. **43**: p. 3-19.
44. Kamarianakis, Y., W. Shen, and L. Wynter, *Real-time road traffic forecasting using regime-switching space-time models and adaptive LASSO.* Applied Stochastic Models in Business and Industry, 2012. **28**(4): p. 297-315.
45. Chan, K.Y., et al., *Prediction of Short-Term Traffic Variables Using Intelligent Swarm-Based Neural Networks.* IEEE Transactions on Control Systems Technology, 2013. **21**(1): p. 263-274.
46. Dia, H., *An object-oriented neural network approach to short-term traffic forecasting.* European Journal of Operational Research, 2001. **131**(2): p. 253-261.
47. Quek, C., M. Pasquier, and B.B.S. Lim, *POP-TRAFFIC: a novel fuzzy neural approach to road traffic analysis and prediction.* IEEE Transactions on Intelligent Transportation Systems, 2006. **7**(2): p. 133-146.
48. Heilmann, B., et al., *Predicting Motorway Traffic Performance by Data Fusion of Local Sensor Data and Electronic Toll Collection Data.* Computer-Aided Civil and Infrastructure Engineering, 2011. **26**(6): p. 451-463.

49. Wang, J. and Q. Shi, *Short-term traffic speed forecasting hybrid model based on Chaos–Wavelet Analysis-Support Vector Machine theory.* Transportation Research Part C: Emerging Technologies, 2013. **27**: p. 219-232.
50. Djuric, N., et al., *Travel Speed Forecasting by Means of Continuous Conditional Random Fields.* Transportation Research Record: Journal of the Transportation Research Board, 2011. **2263**: p. 131-139.
51. Mizuguchi, A., K. Demachi, and M. Uesaka, *Establish of the prediction system of chest skin motion with SSA method.* International Journal of Applied Electromagnetics & Mechanics, 2010. **33**(3/4): p. 1529-1533.
52. Chhatkuli, R.B., et al., *Dynamic Image Prediction Using Principal Component and Multi-Channel Singular Spectral Analysis: A Feasibility Study.* Open Journal of Medical Imaging, 2015. **05**(03): p. 133-142.
53. Hassani, H., S. Heravi, and A. Zhigljavsky, *Forecasting UK Industrial Production with Multivariate Singular Spectrum Analysis.* Journal of Forecasting, 2013. **32**(5): p. 395-408.
54. Hassani, H., et al., *Predicting Global Temperature Anomaly: A Definitive Investigation Using an Ensemble of Twelve Competing Forecasting Models*, in *Department of Economics Working Paper Series*, U.o. Pretoria, Editor. 2015.
55. Alessio, S.M., *Singular Spectrum Analysis (SSA).* 2016: p. 537-571.
56. Patterson, K., et al., *Multivariate singular spectrum analysis for forecasting revisions to real-time data.* Journal of Applied Statistics, 2011. **38**(10): p. 2183-2211.
57. Elsner, J.B. and A.A. Tsonis, *Singular Spectrum Analysis: A New Tool in Time Series Analysis*. 1996.
58. Vitanov, N.K., K. Sakai, and Z.I. Dimitrova, *SSA, PCA, TDPSC, ACFA: Useful combination of methods for analysis of short and nonstationary time series.* Chaos, Solitons & Fractals, 2008. **37**(1): p. 187-202.

59. Cressie, N. and C.K. Wikle, *Statistics for Spatio-temporal Data*. 2011, New Jersey: Wiley.
60. Hassani, H. and D. Thomakos, *A review on singular spectrum analysis for economic and financial time series.* Statistics and Its Interface, 2010. **3**(3): p. 377-397.
61. Asif, M.T., et al. *Data compression techniques for urban traffic data*. in *2013 IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems (CIVTS)*. 2013.
62. Chen, T., E. Martin, and G. Montague, *Robust probabilistic PCA with missing data and contribution analysis for outlier detection.* Computational Statistics & Data Analysis, 2009. **53**(10): p. 3706-3716.
63. Turk, M. and A. Pentland, *Eigenfaces for Recognition.* Journal of Cognitive Neuroscience, 1991. **3**(1): p. 71-86.
64. A.Loskutov, I.Istomin, and O.Kotlyarov, *Data Analysis: Generalisations of the Local Approximation Method by Singular Spectrum Analysis.* ArXiv:nlin/0109022, 2001.
65. Treiber, M., A. Kesting, and R.E. Wilson, *Reconstructing the Traffic State by Fusion of Heterogeneous Data.* Computer-Aided Civil and Infrastructure Engineering, 2011. **26**(6): p. 408-419.
66. Akaike, H. *Information Theory and an Extension of the Maximum Likelihood Principle In B. N. Petrov and F. Csaki (Eds.)*. in *Second international symposium on information theory*. 1973. Budapest: Academiai Kiado.
67. Kirby, H.R., S.M. Watson, and M.S. Dougherty, *Should we use neural networks or statistical models for short-term motorway traffic forecasting?* International Journal of Forecasting, 1997. **13**(1): p. 43-50.
68. TRB, *Highway Capacity Manual 2010*. 2010, Washington, D.C.: Transportation Research Board of the National Academies.

69. Stephanopoulos, G., P.G. Michalopoulos, and G. Stephanopoulos, *Modelling and analysis of traffic queue dynamics at signalized intersections.* Transportation Research Part A: General, 1979. **13**(5): p. 295-307.
70. McLachlan, G.J., *Finite mixture models*, D. Peel, Editor. 2000, New York : Wiley: New York.
71. Webster, F.V., *Traffic Signal Settings*. 1958: H.M. Stationery Office.
72. Michalopoulos, P.G., G. Stephanopoulos, and G. Stephanopoulos, *An application of shock wave theory to traffic signal control.* Transportation Research Part B: Methodological, 1981. **15**(1): p. 35-51.
73. Sharma, A., D. Bullock, and J. Bonneson, *Input-Output and Hybrid Techniques for Real-Time Prediction of Delay and Maximum Queue Length at Signalized Intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2007. **2035**: p. 69-80.
74. Liu, H.X., et al., *Real-time queue length estimation for congested signalized intersections.* Transportation Research Part C: Emerging Technologies, 2009. **17**(4): p. 412-427.
75. Cassidy, M.J. and L.D. Han, *Proposed Model for Predicting Motorist Delays at Two-Lane Highway Work Zones.* Journal of Transportation Engineering, 1993. **119**(1): p. 27-42.
76. Jiang, X. and H. Adeli, *Object-Oriented Model for Freeway Work Zone Capacity and Queue Delay Estimation.* Computer-Aided Civil and Infrastructure Engineering, 2004. **19**(2): p. 144-156.
77. Lighthill, M.J. and G.B. Whitham, *On Kinematic Waves. II. A Theory of Traffic Flow on Long Crowded Roads.* Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 1955. **229**(1178): p. 317.
78. Richards, P.I., *Shock Waves on the Highway.* Operations Research, 1956. **4**(1): p. 42-51.

79. Geroliminis, N. and A. Skabardonis, *Prediction of Arrival Profiles and Queue Lengths Along Signalized Arterials by Using a Markov Decision Process.* Transportation Research Record: Journal of the Transportation Research Board, 2005. **1934**: p. 116-124.
80. Kerner, B.S., *The Physics of Traffic: Empirical Freeway Pattern Features, Engineering Applications, and Theory*. Understanding Complex Systems, ed. J.A.S. Kelso. 2004, Berlin, Heidelberg, New York: Springer.
81. Comert, G. and M. Cetin, *Queue length estimation from probe vehicle location and the impacts of sample size.* European Journal of Operational Research, 2009. **197**(1): p. 196-202.
82. Ban, X., P. Hao, and Z. Sun, *Real time queue length estimation for signalized intersections using travel times from mobile sensors.* Transportation Research Part C: Emerging Technologies, 2011. **19**(6): p. 1133-1156.
83. Bozdogan, H., *Choosing the Number of Component Clusters in the Mixture-Model Using a New Informational Complexity Criterion of the Inverse-Fisher Information Matrix*, in *Information and Classification: Concepts, Methods and Applications Proceedings of the 16th Annual Conference of the "Gesellschaft für Klassifikation e.V." University of Dortmund, April 1–3, 1992*, O. Opitz, B. Lausen, and R. Klar, Editors. 1993, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 40-54.
84. Schwarz, G., *Estimating the Dimension of a Model.* Ann. Statist., 1978. **6**(2): p. 461-464.
85. Celeux, G. and G. Soromenho, *An entropy criterion for assessing the number of clusters in a mixture model.* Journal of Classification, 1996. **13**(2): p. 195-212.
86. Edie, L.C., *Car-Following and Steady-State Theory for Noncongested Traffic.* Operations Research, 1961. **9**(1): p. 66-76.

87. Daganzo, C.F., *A behavioral theory of multi-lane traffic flow. Part I: Long homogeneous freeway sections.* Transportation Research Part B: Methodological, 2002. **36**(2): p. 131-158.
88. Treiber, M. and D. Helbing, *Reconstructing the spatio-temporal traffic dynamics from stationary detector data.* Cooper@tive Tr@nsport@tion Dyn@mics, 2002. **1**: p. 3.1–3.24.
89. Treiber, M. and A. Kesting, *Traffic Flow Dynamics: Data, Models and Simulation*. 2012, Heidelberg New York Dordrecht London: Springer.
90. FHWA, *Manual on Uniform Traffic Control Devices for Streets and Highways*. 2009 ed. 2009, Washington, D.C.: Federal Highway Administration, U.S. Department of Transportation.
91. Han, L., J.-M. Li, and T. Urbanik, *Control-Type Selection at Isolated Intersections Based on Control Delay Under Various Demand Levels.* Transportation Research Record: Journal of the Transportation Research Board, 2008. **2071**: p. 109-116.
92. Rodegerdts, L., et al., *Roundabouts: An Informational Guide*, in *NCHRP Report*. 2010: Washington, D.C.
93. TRB, *Highway Capacity Manual 2000*. 2000, Washington, D.C.: Transportation Research Board National Research Council.
94. Fambro, D. and N. Rouphail, *Generalized Delay Model for Signalized Intersections and Arterial Streets.* Transportation Research Record: Journal of the Transportation Research Board, 1997. **1572**: p. 112-121.
95. Benekohal, R.F. and Y.M. El-Zohairy, *Multi-regime arrival rate uniform delay models for signalized intersections.* Transportation Research Part A: Policy and Practice, 2001. **35**(7): p. 625-667.
96. Sazi Murat, Y., *Comparison of fuzzy logic and artificial neural networks approaches in vehicle delay modeling.* Transportation Research Part C: Emerging Technologies, 2006. **14**(5): p. 316-334.

97. Tian, Z., et al., *Variations in capacity and delay estimates from microscopic traffic simulation models.* Transportation Research Record, 2002. **1802**: p. 23-31.
98. Khatib, Z. and M. Kyte. *Framework to consider the effect of uncertainty in forecasting the level of service of signalized and unsignalized intersections*. in *Fourth International Symposium on Highway Capacity*. 2000. Maui, HI: Transportation Research Board.
99. Tarko, A.P. and M. Tracz. *Uncertainty in saturation flow predictions*. in *Fourth International Symposium on Highway Capacity*. 2000. Maui, HI: Transportation Research Board.
100. Kimber, R.M., M. Mcdonald, and N.B. Hounsell, *The prediction of saturation flows for single road junctions controlled by traffic signals*, in *Transport Research Laboratory Report*. 1986: Wokingham, GB.
101. Daganzo, C.F., *Increasing Model Precision Can Reduce Accuracy.* Transportation Science, 1987. **21**(2): p. 100-105.

APPENDIX

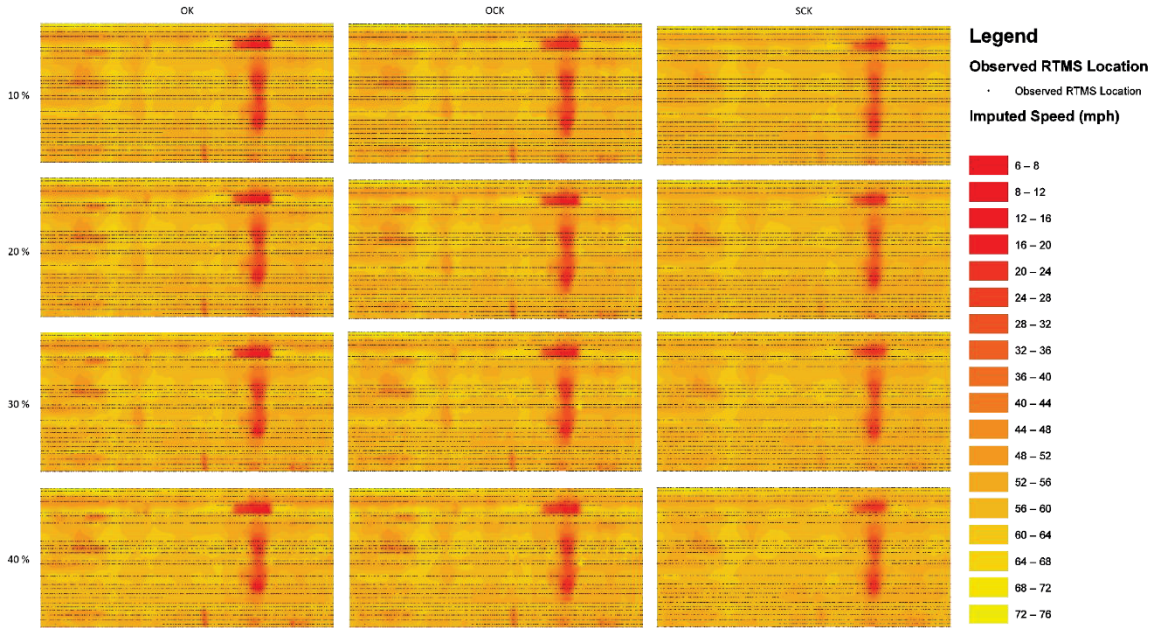


Figure A-1 MCAR patterns and imputed speed.

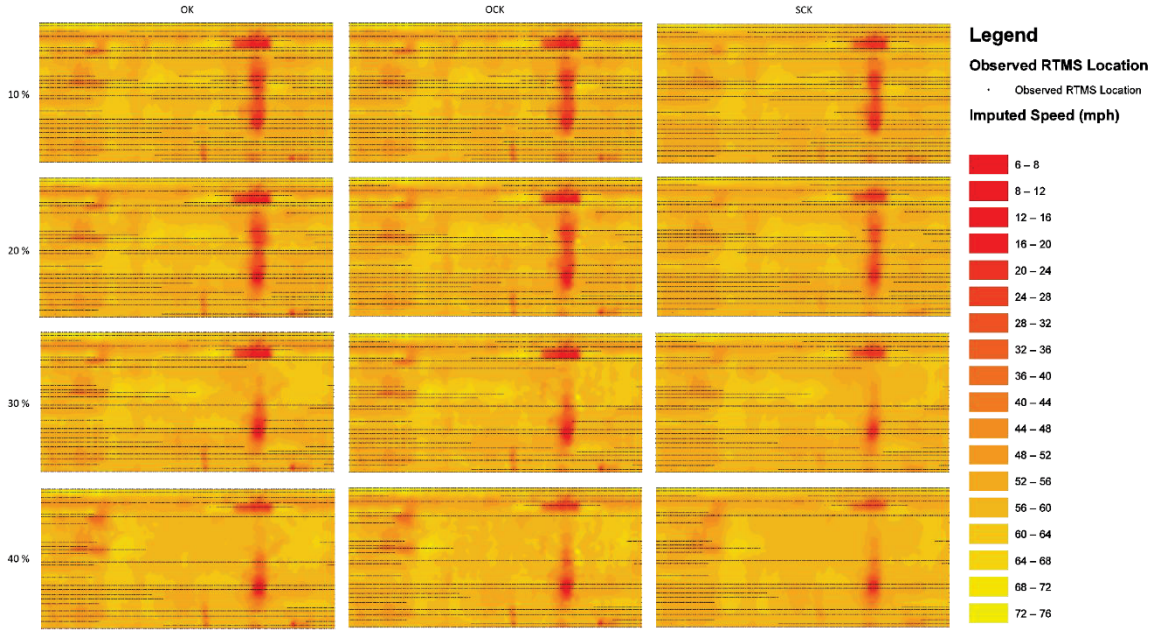


Figure A-2 MAR patterns and imputed speed.

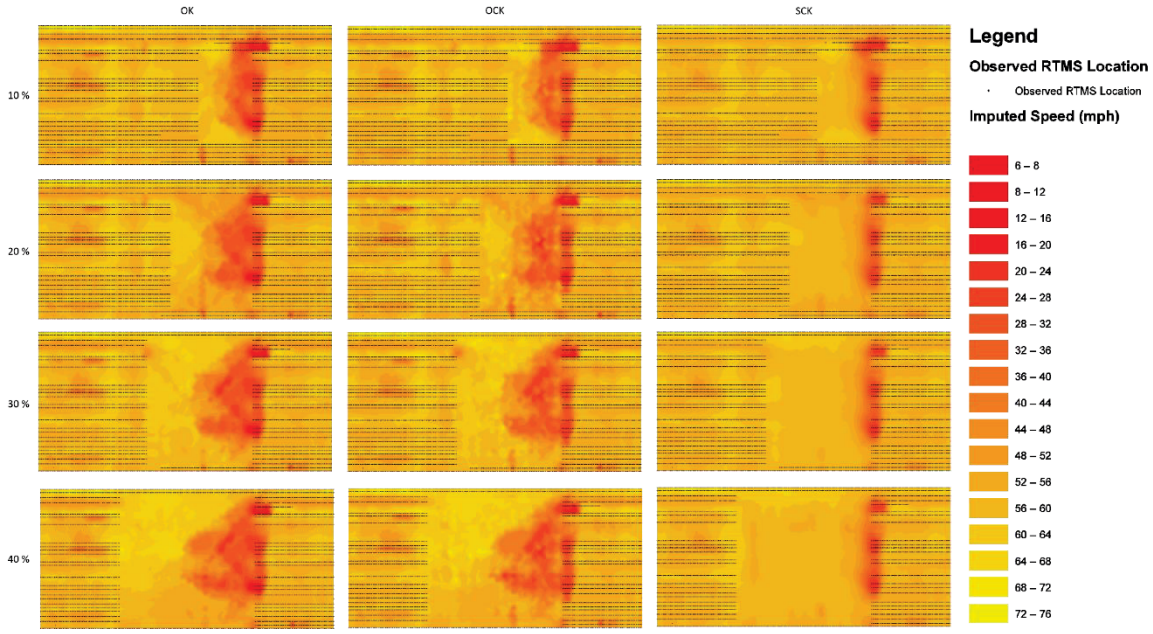


Figure A-3 MNAR patterns and imputed speed.

Table A-1 Summary of imputation errors.

Measurement	Missing Pattern	Missing Rate	OK	OCK	SCK
MAE (mile/hour)	MCAR	10%	1.981	2.475	3.071
		20%	1.939	2.320	3.019
		30%	1.933	2.453	3.179
		40%	2.033	2.650	3.216
	MAR	10%	4.915	5.082	5.118
		20%	5.502	5.477	5.750
		30%	5.476	5.409	5.495
		40%	5.528	5.663	5.475
	MNAR	10%	11.393	10.819	6.507
		20%	11.256	11.354	5.754
		30%	9.667	9.370	4.916
		40%	9.163	8.867	4.658
MAPE (%)	MCAR	10%	4.949	5.860	7.472
		20%	4.724	5.335	7.171
		30%	4.592	5.591	7.647
		40%	4.647	5.865	7.565
	MAR	10%	8.628	8.747	9.017
		20%	10.620	10.381	11.316
		30%	12.338	11.870	12.318
		40%	13.151	13.139	13.206
	MNAR	10%	20.636	19.488	12.449
		20%	20.539	20.902	11.595
		30%	17.711	17.140	10.105
		40%	16.648	16.066	9.296

VITA

Bumjoon Bae was born in Daegu, South Korea. In 2009, he graduated from Ajou University with a Bachelor's degree in Transportation System Engineering with an honor. In 2011, Bumjoon graduated from Seoul National University with his Master's degree in City Planning in Transportation Studies. From 2011 to 2014, Bumjoon worked as a transportation engineer at Korea Road Traffic Authority, a government agency in transportation safety. In 2017, he was granted a doctoral degree in Civil Engineering with concentration in Transportation Engineering and Master's degree in Statistics at the University of Tennessee, Knoxville (UT).

As a Ph.D. student, Bumjoon worked on multiple research projects in traffic operations, including virtual transportation management center establishment, travel time reliability analysis, traffic queue detection and prediction, and so on. His research interests include urban traffic operations, machine learning applications in transportation studies, spatio-temporal traffic analysis, short-term traffic forecasting, and traffic flow theory related to connected and autonomous vehicles.

During his graduate studies, Bumjoon received several scholarships and awards, including Intelligent Transportation Society of Tennessee's annual scholarship award, William L. Moore, Jr. scholarship award from Tennessee Section Institute of Transportation Engineers (TSITE), TSITE student paper competition award, Frank Richter AARS scholarship from American Association of Railroad Superintendents (first recipient in UT), Graduate Student Senate travel awards, and so on.