USE OF MULTISENSOR DATA IN MODELING FREEWAY TRAVEL TIME: VARIABILITY ANALYSIS AND PREDICTION

by

Zhen Chen

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Infrastructure and Environmental Systems

Charlotte

2019

Approved by:

____

Dr. Wei Fan

____

Dr. Martin Kane

____

Dr. Miguel Pando

____

Dr. David Weggel

____

Dr. Jay Wu

____

Dr. Yu Wang

ABSTRACT

ZHEN CHEN. Use of multisensor data in modeling freeway travel time: variability analysis and prediction. (Under the direction of DR. WEI FAN)

Nowadays anonymous vehicle probe data have been greatly improved in both data coverage and data fidelity. Thus, vehicle probe data have become a reliable source for freeway travel time analysis. The travel time variability is highly complex as it is affected by a wide variety of factors. A better understanding of travel time variability patterns can help the decision makers plan, design, operate, and manage a more efficient highway system.

Moreover, travel time prediction also plays a significant role in traffic data analysis and applications as it can assist in route planning and reducing traffic congestion. With the development of artificial intelligence technologies, various novel prediction methods have been developed accordingly in recent years. Machine learning is an example of a data driven method which aims to increase efficiency and accuracy of predictions. Recently, different machine learning-based approaches, such as neural network, ensemble learning, and support vector machines (SVM), have been employed by the researchers and the results indicate that such approaches for prediction are adaptable and can give better performances than traditional models.

This research is intended to systematically analyze how travel time distributes and varies with respect to the time of day, day of week, year, and weather conditions. In addition, an advanced machine learning-based approach (i.e. XGBoost model) is employed to predict the freeway travel time. Detailed information about the input variables and data pre-processing is presented. Parameters of the XGBoost model are introduced and the parameter tuning process is also discussed. The relative importance of each variable in the model is presented and interpreted. Optimized modeling results of the proposed XGBoost travel time prediction model are evaluated and compared with those of the gradient boosting model. The results also demonstrate that the developed XGBoost travel time prediction model significantly improves the computation accuracy and efficiency. Summary and conclusions of the whole study are made and further research directions are given at the end of study.

ACKNOWLEDGEMENTS

Firstly, I would like to express my deepest appreciation to my advisor Professor Wei Fan for the continuous support of my Ph.D. study and related research. I'm grateful for his patience, motivation, and immense knowledge. His guidance helped me in researching and writing this dissertation.

I would also like to thank the rest of my thesis committee: Professor Martin Kane, Professor Miguel Pando, Professor David Weggel, Professor Jay Wu, and Professor Yu Wang, for their insightful comments and questions on this dissertation which inspired me to widen my research from various perspectives.

I am grateful to my groupmates in CAMMSE lab for their help in my life and study. I would also like to thank Professor Mark Magalotti, who taught me a lot when I worked with him in University of Pittsburgh.

Last but not the least, I would like to thank my family and friends: my father Shuming Chen and my mother Yuhong Peng for supporting me spiritually throughout writing this thesis and my life in general. I am also grateful to my friends, Hannah, Josh, BV and Jiawei, who have given me endless support in the U.S.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AADT	annual average daily traffic
AASHTO	American association of state highway and transportation officials
AD	Anderson-Darling
ANPR	automated number plate recognition
ARIMA	autoregressive integrated moving average
ARMA	average moving average
AVI	automatic vehicle identification
BI	buffer index
BP	back propagation
BPR	bureau of public roads
C&RT	classification and regression tree
CV	coefficient of variation
DOT	department of transportation
DOW	day of week
FHWA	federal highway administration
FOC	frequency of congestion
FOSIM	freeway operations simulation
GARCH	generalized autoregressive conditional heteroscedasticity
GPD	generalized Pareto distribution
KF	Kalman filter
K-NN	k-nearest neighbor

LSTM	long short-term memory
MAC	media access control
MAE	mean absolute error
MAPE	mean absolute percentage error
MFD	macroscopic fundamental diagram
MI	misery index
MSE	mean squared error
NCHRP	national cooperative highway research program
OD	origin-destination
PEMS	performance measurement system
PTI	planning time index
RITIS	regional integrated transportation information system
RMSE	root mean squared error
SHRP 2	strategic highway research program 2
SVM	support vector machines
TMC	traffic message channel
TOD	time of day
TTR	travel time reliability
TTSD	travel time standard deviation
XGBoost	extreme gradient boosting

CHAPTER 1: INTRODUCTION

1.1. Problem Statement and Motivation

Nowadays anonymous vehicle probe data have been greatly improved in both data coverage and data fidelity, and thus have become a reliable source for freeway travel time analysis. The travel time variability is highly complex as it is affected by a wide variety of factors. These include aspects such as time of day (TOD), day of week (DOW), segment locations, and weather conditions. A better understanding of travel time variability patterns can greatly help the decision makers plan, design, operate, and manage a more efficient highway system.

Moreover, travel time prediction also plays a significant role in traffic data analysis and applications as it can greatly help in route planning and reducing traffic congestion. Traditionally, the methods such as linear regression and time series models have been widely applied to predict travel times using historical data. However, with the consideration of effectiveness, accuracy, and feasibility, these models may become outdated and replaceable. With the development of artificial intelligence technologies, various novel prediction methods have been developed accordingly in recent years. Machine learning is an example of a data driven method which aims to increase efficiency and accuracy of the prediction. Recently, different machine learning-based approaches, such as neural network, ensemble learning, and support vector machines (SVM), have been employed by the researchers and the results indicate that such approaches for prediction are adaptable and can give better performances than traditional models. Therefore, the machine learning-based approach is selected for the travel time prediction in this study.

This research is intended to systematically analyze how travel time distributes and varies with respect to the TOD, DOW, year, and weather conditions. In addition, an advanced machine learning-based approach (i.e. XGBoost model) is employed to predict the freeway travel time.

1.2. Study Objectives

The proposed work in this research is intended to fulfill the following objectives:

1. To select the most appropriate travel time reliability (TTR) measure that could properly describe travel time variability, such as planning time index. Typical segments based on historical TTR ratings are also selected to illustrate the characteristics in different cases.
2. To systematically analyze the travel time variability patterns with the consideration of time of day, day of week, year, and weather. The potential reasons of the variability patterns are analyzed.
3. To develop the travel time prediction model using an advanced, efficient and accurate machine learning-based approach.
4. To examine and evaluate the developed prediction models on a real-world freeway so that the gaps between the theoretical research and the application of the developed travel time prediction model can be bridged.

1.3. Expected Contributions

In order to better understand freeway travel time characteristics, travel time variability patterns under different conditions are studied in this research with the help of TTR measures. The validation of machine learning-based travel time prediction model is also presented. The expected contributions from this research are summarized as follows:

1. Ability to select appropriate TTR measures to analyze travel time variability and identify typical segments for further analysis with the help of such TTR measures.
2. Ability to understand the travel time variability characteristics of selected segments with the consideration of time of day, day of week, year and weather.
3. Ability to develop an advanced, efficient and accurate travel time prediction model.
4. Ability to predict real world freeway travel time by using the developed prediction model.

1.4. Research Overview

Figure 1.1 shows the research structure. In Chapter 1, the significance and motivation of the travel time variability analysis and prediction have been discussed, followed by the description of study objectives and expected contributions.

Chapter 2 presents a comprehensive review of the studies related to travel time variability analysis and prediction. Previous approaches that were adopted to analyze travel time variability are classified into four categories: (1) basic travel time variability analysis; (2) network level travel time variability analysis; (3) travel time variability analysis considering weather conditions and incidents; (4) travel time variability analysis considering multiple influencing factors. In addition, the machine learning-based travel time prediction methodologies are reviewed and summarized in this chapter. In detail, methods used by the reviewed studies including the neural network approach, ensemble learning approach, K-nearest neighbor (K-NN) approach, and support vector machine approach, will be presented.

Chapter 3 describes the basic information needed to analyze travel time variability, including the travel time data and historical weather data utilized in this study. Detailed information about the raw travel time data source is described first, followed by the discussions about weather data collection. The data processing steps are also explained in detail in this chapter.

Chapter 4 discusses the analysis of travel time variability patterns. The study location identification process based on the selected TTR measure is described first. With the help of planning time index (PTI), travel time variability patterns of the selected segments under all conditions (including the DOW and different weather conditions) are also described in detail.

Chapter 5 presents the travel time prediction methodology which is utilized in this study. The idea of ensemble learning is introduced first. Detailed information on the decision tree algorithm, bagging algorithm, and boosting algorithm is presented. The basic information about the Random Forest and gradient boosting models is described including advantages and disadvantages. An introduction of the XGBoost model is also presented in this chapter. Advantages of the XGBoost model are listed. The detailed process of the XGBoost model is described including the objective function, regularization terms, and model score.

Chapter 6 discusses the validation steps of the proposed XGBoost-based travel time prediction model based on the data described in Chapter 3. Selected features include, but are not limit to, the following: TOD, DOW, month of year, year, weather conditions, segment characteristics, etc. Detailed information about the input variables and data pre-processing is presented. The parameters of the XGBoost model are introduced and the parameter tuning process is also discussed. The experiment results could give a clear picture of how the analyzed parameters impact the prediction performance.

Chapter 7 presents the interpretation and evaluation of the numerical results of the developed XGBoost model. The relative importance of each variable in the model is presented and interpreted. In order to examine the accuracy and effectiveness of the proposed model, this chapter also evaluates the optimized modeling results of the proposed XGBoost travel time prediction model and compares them with those of the gradient boosting model. The results also demonstrate that the developed XGBoost travel time prediction model significantly improves the computation accuracy and efficiency.

Chapter 8 concludes the study with a summary of the discussions about the travel time variability analysis, the developed travel time prediction model, and the modeling results. Future research possibilities are also provided in this chapter.

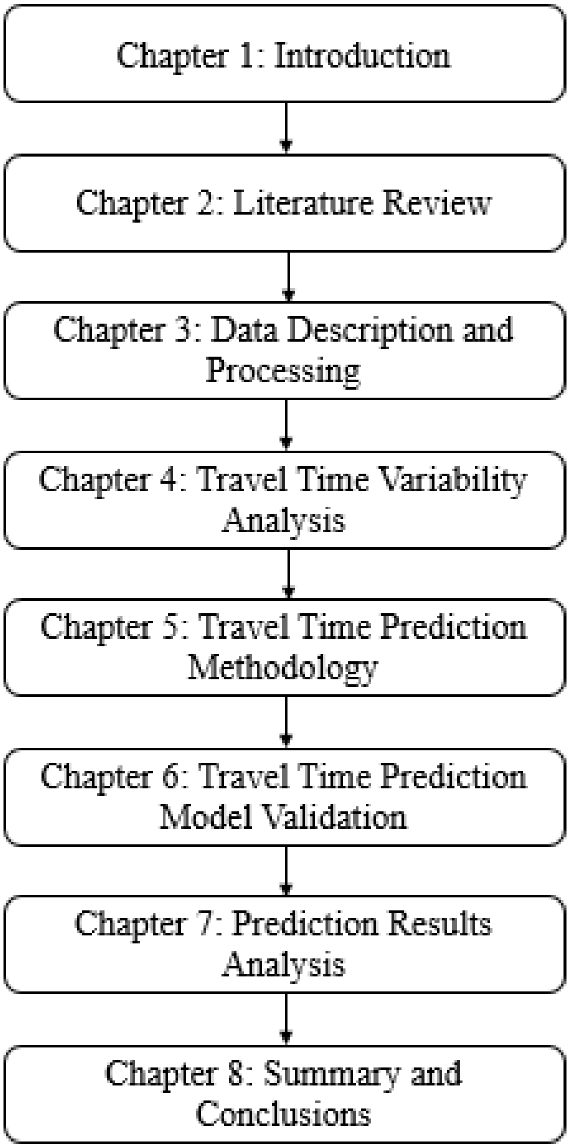

Chapter 1: Introduction
Chapter 2: Literature Review
Chapter 3: Data Description and Processing
Chapter 4: Travel Time Variability Analysis
Chapter 5: Travel Time Prediction Methodology
Chapter 6: Travel Time Prediction Model Validation
Chapter 7: Prediction Results Analysis
Chapter 8: Summary and Conclusions

FIGURE 1.1: Research structure

CHAPTER 2: LITERATURE REVIEW

2.1. Introduction

This chapter provides a comprehensive review of travel time studies regarding TTR definitions, existing TTR measures, travel time variability modeling and analysis, and travel time prediction methodologies, etc. This should give a clear picture of existing efforts toward the modeling of travel time variability and travel time prediction.

The following sections are organized as follows. Section 2.2 presents several definitions of TTR, followed by a list of TTR measures. Section 2.3 gives a comprehensive review of existing methods of travel time variability analysis, which include travel time distribution-based studies, network level travel time variability studies, travel time variability analysis with the consideration of incidents/weather studies and travel time variability analysis with the consideration of multiple influencing factors. Section 2.4 presents several common travel time prediction methods. Finally, section 2.5 concludes this chapter with a summary.

2.2. Travel Time Variability Analysis Theoretical Background

2.2.1. Definitions of Travel Time Reliability

In order to learn the theoretical background of travel time variability analysis, the concept of travel time reliability (TTR) is briefly introduced in this section first. Typically, the TTR can be used as a measure of service to describe the variability of travel time (Chen et al. 2003).

Different definitions of travel time reliability have been developed in different studies. It will be helpful to review the existing definitions in different studies to clarify the concept of travel time reliability and its measurement. This section briefly reviews existing 'reliability' and 'travel time reliability' definitions. Table 2.1 provides a summary of existing travel time reliability definitions in chronological order.

Charles (1997) defined reliability as "*the probability that a component or system will perform a required function for a given period of time when used under stated operating conditions. It is the probability of a non-failure over time*." This definition is similar to the other definitions used in reliability engineering (Elefteriadou and Cui, 2007).

In the transportation area, different kinds of reliability definitions have developed by previous studies. The definitions include system reliability, travel time reliability and network reliability. Turner et al. (1996) defined trip time reliability as the range of travel times experienced during a large number of daily trips. The National Cooperative Highway Research Program (NCHRP) report 398 (1997) defined travel time reliability as "*the impact of non-recurrent congestion on the transportation system*." In the NCHRP report 399 (1998), travel time reliability was defined as "*a measure of the variability of travel time*". California Transportation Plan (1998) defined reliability as "the level of variability between the expected travel time and the actual travel time experienced." Florida Department of Transportation (DOT) (2011) defined the highway travel time reliability as "*the percent of travel that takes no longer than the expected travel time plus a certain acceptable additional time*." They also defined three major components of reliability: travel time, expected travel time, and acceptable additional time. The American Association of State Highway and Transportation Officials (AASHTO)'s freight report (2002) defined reliability as "*the percent of on-time performance for a given time schedule*", and this definition was provided for freight transportation. Recker et al. (2004) defined both path and Origin-Destination (OD) travel time reliability. Specifically, the path travel time reliability was defined as "*the probability that the travel time of a given path is within an acceptable threshold*" and the OD travel time reliability was defined as "*the probability that the weighted average travel time of a given OD pair is within an acceptable threshold*." The Federal Highway Administration (FHWA) (2012) gave a formal definition of travel time reliability, which is: "*the consistency or dependability in travel times, as measured from day-to-day and/or across different times of the day*." The Strategic Highway Research Program 2 (SHRP 2) Project (2014) defined travel time reliability as "*the variability in travel times that occur on a facility or for a trip over the course of time; and the number of times (trips) that either "fail" or "succeed" in accordance with a predetermined performance standard or schedule*."

TABLE 2.1: Summary of existing travel time reliability definitions

Author/Agency	Year	Reliability/Travel Time Reliability Definition
Turner et al.	1996	The range of travel times experienced during a large number of daily trips.
Charles	1997	The probability that a component or system will perform a required function for a given period of time when used under stated operating conditions. It is the probability of a non-failure over time.
NCHRP Report 398	1997	The impact of non-recurrent congestion on the transportation system.
NCHRP Report 399	1998	A measure of the variability of travel time.
California Transportation Plan	1998	The level of variability between the expected travel time and the actual travel time experienced.
AASHTO's Freight Report	2002	The percent of on-time performance for a given time schedule.
Elefteriadou and Cui	2007	The probability of a device performing its purpose adequately for the period of time intended under the stated operating conditions.
Florida DOT	2011	The percent of travel that takes no longer than the expected travel time plus a certain acceptable additional time.
FHWA	2012	The consistency or dependability in travel times, as measured from day-to-day and/or across different times of the day
Vandervalk et al. (SHRP 2 project)	2014	The variability in travel times that occur on a facility or for a trip over the course of time; and the number of times (trips) that either "fail" or "succeed" in accordance with a predetermined performance standard or schedule

2.2.2. Travel Time Reliability Measures

This section introduces the characteristics of different travel time reliability measures. Table 2.2 provides a summary of the TTR measures discussed in this section in chronological order.

2.2.2.1. Standard deviation

Standard deviation is a classical statistical measure and usually used as a proxy for other reliability measures (Charles, 1997). However, the use of standard deviation as a reliability performance measure was discouraged by some studies (USDOT, 1996 and NCHRP Report 618, 2008) because "*it is not easily understood by nontechnical audiences nor easily related to everyday commuting experiences, and it treats early and late arrivals with equal weight, whereas the public cares much more about late arrival.*"

2.2.2.2. Coefficient of variation (CV)

The average travel time and standard deviation values can be combined and used to generate a value which is called coefficient of variation (CV). The CV is calculated as the ratio of the standard deviation to the mean. The use of CV is also discouraged by some studies with the same concern as about the usage of standard deviation. However, it still being utilized by some researchers.

$$\text{CV} = \frac{\text{Standard deviation}}{\text{Average travel time}}$$

2.2.2.3. Percent variation

The average travel time and standard deviation values can also be combined as a ratio to produce a value, which was recommended by the 1998 California Transportation Plan.

$$\text{Percent variation} = \frac{\text{Standard deviation}}{\text{Average travel time}} \times 100\% = \text{CV} \times 100\%$$

This measure has the same mathematical characteristics as the CV. However, it is easier for the public to understand percent variation as it is expressed as a percentage of average travel time. This measure was adopted by the 1998 California Transportation Plan (1998) and recommended by Lomax et al. (1997) and NCHRP Report 618 (2008).

2.2.2.4. Variability index

The variability index is a ratio of *peak to off-peak variation in travel conditions*. The index is calculated as *"a ratio of the difference in the upper and lower 95% confidence intervals between the peak period and the off-peak period"* (Lomax et al., 1997 and Florida DOT, 2011).

$$\text{Variability Index} = \frac{\text{Difference in peak period confidence intervals}}{\text{Difference in off peak period confidence intervals}}$$

Because the interval differences in the off-peak periods are usually lower than the differences in the peak period, the value of variability index is usually greater than 1.

2.2.2.5. 90th/95th percentile travel times

90th/95th percentile travel times are both basic TTR measures which have been widely used in the world. These indexes indicate how much delay will be on the heaviest travel days and were introduced as one of the four recommended travel time reliability measures by FHWA. The 90th or 95th percentile travel times are "*usually reported in minutes and seconds, they could be easily understood by roadway users who are familiar with their trips*."

However, the disadvantage of this measure is "*not being easily compared across trips with the consideration of different trip lengths*."

2.2.2.6. Buffer index (BI)

Buffer index (BI) represents the extra time required by the travelers to arrive on time in addition to the travel time under average conditions. It was introduced as one of the four recommended measures by FHWA. The BI is computed as "*the difference between the 95th percentile travel time and average travel time, divided by the average travel time*" (Lomax et al., 1997). The equation is shown below:

$$\mathrm{BI} = \frac{\text{95th percentile time} - \text{average travel time}}{\text{average travel time}} \times 100\%$$

2.2.2.7. Planning time index (PTI)

Planning time index (PTI) was also introduced as one of the four recommended measures by FHWA. It represents *the total time needed to plan for an on-time arrival 95% of the time (total travel time that should be planned when an adequate buffer time is included), computed as 95th percentile travel time divided by free-flow travel time.* The equation of PTI is presented below:

$$\mathrm{PTI} = \frac{\text{95th percentile travel time}}{\text{free flow travel time}}$$

The PTI differs from the BI in that it compares near-worst case travel time with that under free-flow traffic condition. For example, a PTI of 1.50 means that, for a 20-minute trip under light traffic condition, the total time that should be planned for the trip is 30 minutes (20 minutes × 1.50 = 30 minutes). PTI is a useful measure as it can be directly combined and used with the travel time index.

2.2.2.8. Frequency of congestion (FOC)

Frequency of congestion (FOC) is a measure introduced as one of the four recommended travel time reliability measures by FHWA, which represents the frequency of congestion exceeding some expected threshold. It can be typically expressed as the percent of days/time that travel times exceed a time threshold *x* or travel speeds fall below a speed threshold *y*. The FOC is relatively easy to compute if continuous traffic data are available, and it is typically reported on weekdays during peak traffic periods.

2.2.2.9. Skew of travel time distribution

The skew statistics is a robust measure introduced by Van Lint and Van Zuylen (2005). It is defined as "*the ratio of the difference between the 90th percentile travel time and the median and the difference between the median and the 10th percentile travel time*." The equation is given below:

$$\lambda^{skew} = \frac{\mathrm{T90} - \mathrm{T50}}{\mathrm{T50} - \mathrm{T10}}$$

2.2.2.10. Width of travel time distribution

The width statistics is a robust measure introduced by Van Lint and Van Zuylen (2005). It is defined as the ratio of the difference between the 90th percentile travel time and the 10th percentile travel time and median travel time. The equation is shown below:

$$\lambda^{width} = \frac{\mathrm{T90} - \mathrm{T10}}{\mathrm{T50}}$$

2.2.2.11. Misery index

Misery Index is a measure that can indicate the length of delay of only the worst trips. It is usually computed by subtracting the average travel rate from the upper 20 percent of travel rates. This yields the time difference between the average trip and the slowest 20 percent of trips. The equation is below:

$$\text{Misery index} = \frac{\text{Average travel rate (Top 20\% trips)}}{\text{Average travel rate}} - 1$$

TABLE 2.2: Summary of travel time reliability measures

Measure	Author/Agency	Equation
Standard deviation	Dowling et al. (2009); Pu (2011)	Standard deviation
Coefficient of variation	Pu (2011)	$\text{Coefficient variation} = \frac{\text{Standard deviation}}{\text{Average travel time}}$
Present variation	1998 California Transportation Plan; Lomax et al. (1997); NCHRP Report 618 (2008)	$\text{Percent variation} = \frac{\text{Standard deviation}}{\text{Average travel time}} \times 100\%$
Variability Index	Lomax et al. (1997); Albert (2000)	$\frac{\text{Difference in peak period confidence intervals}}{\text{Difference in off peak period confidence interval}}$
90th/95th Percentile Travel Times:	FHWA	90th/95th Percentile Travel Times
Buffer Index	FHWA	$\frac{\text{95th precentile time} - \text{average travel time}}{\text{average travel time}} \times 100\%$
Planning Time Index	FHWA	$\frac{\text{95th percentile travel time}}{\text{free flow travel time}}$
Frequency of Congestion	FHWA	Frequency of trips exceeding a threshold value
Skew of travel time distribution	Van Lint and Van Zuylen (2005)	$\lambda^{skew} = \frac{\text{T90} - \text{T50}}{\text{T50} - \text{T10}}$
Width of travel time distribution	Van Lint and Van Zuylen (2005)	$\lambda^{width} = \frac{\text{T90} - \text{T10}}{\text{T50}}$
Misery Index	Lomax et al. (1997)	$\text{Misery index} = \frac{\text{Average travel rate (Top 20\% trips)}}{\text{Average travel rate}} - 1$

2.3. Travel Time Variability Analysis Methods

Basically, the travel time variability pattern can be analyzed based on travel time distribution data only. However, to investigate the impacts of nonrecurring congestion, different sources of travel time variability including traffic incidents, inclement weather, and work zones were also studied by different researchers around the world. This section reviews these studies by classifying previous studies into 5 categories including basic travel time variability analysis studies; Network travel time variability studies; Travel time variability studies with the consideration of weather impact; Travel time variability studies with the consideration of incident impact, and travel time variability studies with the consideration of multiple influencing factors.

2.3.1. Basic Travel Time Variability Analysis

Research studies that used basic data to model travel time variability using TTR measures are reviewed in this section. Table 2.3 provides a summary of the studies reviewed in this section in chronological order.

2.3.1.1. Van Lint and Van Zuylen's research work

Van Lint and Van Zuylen (2005) derived two time-reliability-metrics (skew and width) based on the 90th, 50th and 10th percentile of the day-to-day travel time data. Both metrics can make a clear distinction between different traffic flow conditions (congestion, free or transient). They could also identify the travel time reliability and congestion during a given TOD and DOW time period. The results could be used in discrete choice models and for travel time unreliability visualization on the map.

2.3.1.2. Saberi and Bertini's research work

Saberi and Bertini (2010) prioritized freeway segments with the help of TTR measures based on the archived loop detector data from the Interstate-5 freeway (24 miles long) in Portland, Oregon in the U.S. Several reliability measures were selected and examined using differential reliability maps and compared with travel-time-based measures. The authors found that "*the buffer time index and the coefficient of variance were the most consistent among the measures of reliability*." Their research also showed that freeway segment correlations have high impacts on the variability of corridor travel time and should not be ignored. It was also found that different reliability measures presented different portraits of the reliability aspects on a freeway corridor. However, other factors contributing to the unreliability of travel times were not identified in this research study.

2.3.1.3. Yazici et al.'s research work

Yazici et al. (2012) developed a method to analyze TTR based on DOW and TOD patterns by utilizing GPS data collected from taxis in the New York City. The authors selected coefficient of variation (CV), skewness (λ_{skew}), and width of the distribution (λ_{var}) as the TTR measures and used the Classification and Regression Tree (C&RT) model for the determination of DOW-TOD periods for each selected TTR measure.

The results of the study showed that TTR exhibited time-varying patterns which could be identified during different DOW-TOD periods. Based on the analysis results, the authors found that the *"levels of reliability at the calculated periods generally did not agree well"*, which means that a reliable period identified based on one measure could be found to be an unreliable period using a different measure.

2.3.1.4. Eliasson's research work

Eliasson (2007) used data from the Stockholm's automatic camera system and developed a model for estimating travel time variability in terms of the mean travel time, length of link, and free flow travel time.

The author identified "*a stable relationship between the relative standard deviation of travel time (standard deviation divided by travel time) and the relative increase in travel time (travel time divided by free-flow travel time)*" and then estimated a function to predict how changes in congestion impact the TTR.

The author also investigated the relationship between travel time distribution and different TOD periods. The result showed that *"travel times are approximately normally distributed"* under severe congestion condition. However, the travel time distribution was skewed under low levels of congestion condition.

2.3.1.5. Emam and Ai-Deek's research work

Emam and Ai-Deek (2006) defined reliability as "*the probability that an entity will perform its intended function(s) satisfactorily or without failure for a specified length of time under the stated operating conditions at a given level of confidence*". Based on such definition, the TTR was expressed mathematically using the failure rate (hazard) function. Four different travel time distributions were tested in this study including Weibull, exponential, log-normal, and normal distribution. The Anderson-Darling (AD) goodness-of-fit statistics and error percentages were employed to evaluate model performances. As a result, the log-normal distribution provided the best model fit and was then used to predict TTR of freeway corridors. The proposed methodology was applied to estimate travel time reliability on the I-4 corridor in Orlando, Florida using real-world transportation data collected by dual-loop detectors.

The results indicated that it was more efficient to use the same day of the week (e.g., Mondays) in the estimation of TTR for a roadway segment than to use mixed data (i.e., data collected across multiple weekdays), because of the significant differences between traffic patterns across multiple weekdays. In addition, the researchers also noticed that the new reliability estimation method showed higher sensitivity to geographical locations, which reflects the congestion level and bottlenecks.

2.3.1.6. Sohn and Kim’s research work

Sohn and Kim (2009) presented a method for predicting the dynamic variance in estimating link travel times. The authors adopted the autoregressive moving average-generalized autoregressive conditional heteroscedasticity (ARMA-GARCH) model and employed the generalized Pareto distribution (GPD) in the model to solve the problem of *asymmetry in travel time distribution*.

The authors also used the travel time data which were obtained from the beacon-based probing system in Seoul and performed single and multiperiod predictions. The 90th, 95th, and 99th percentiles of travel times were selected as the TTR measures.

The analysis results showed that the ARMA-GARCH-GPD model was the most promising model for the first four sites. For the other sites without GPD, the ARMA-GARCH was good enough to obtain promising results.

2.3.1.7. Hainen et al.’s research work

Hainen et al. (2011) conducted a study to compute travel time based on the data collected from Bluetooth devices. To examine the impact of bridge closure in Indiana, US, the authors used data from media access control (MAC) addresses from Bluetooth-enabled devices to conduct travel time plots and identify congestion choke points. The authors also estimated the distribution of travel times on four alternate routes. The 25th and the 75th percentile travel times were used as the TTR measures to evaluate the effects of each choice.

This study indicated how to evaluate different route choice based on data collected from Bluetooth devices, sampling methodology and travel time reliability data.

2.3.1.8. Zheng et al.'s research work

Zheng et al. (2016) utilized the data from Automated Number Plate Recognition (ANPR) cameras to study TTR on a corridor in Changsha, China. Two reliability measures (standard deviation and the skewness of travel time) were derived from the travel time distribution model. The authors also investigated the relationship between these two measures and the expected travel time to show the effects of changing travel states. The results showed that the linear relationship could be developed between Travel Time Standard Deviation (TTSD) and mean travel time and skewness.

However, the linear relationships between TTSD and the mean travel time and skewness were not same under different links/days. The regression parameters for a link also linearly depend on the link length.

TABLE 2.3: Summary of basic travel time variability analysis

Year	Author	Location	Data Aggregation	Data Source	Study Periods	TTR Measure(s)	Modeling Algorithm
2005	Van Lint and Van Zuylen	Rotterdam, Netherlands	15-min	N/A	6 a.m.- 8 p.m.	λ^{skew}, λ^{width}	Piecewise linear speed–based trajectory algorithm
2007	Eliasson	Stockholm	15-min	Camera detectors	6:30 a.m.- 8:30 p.m.	Standard deviation	N/A
2006	Emam and Ai-Deek	Orlando, FL, US	5-min	RTMC	3:30 – 6:30 p.m.	Buffer index, Coefficient of variation	Weibull, exponential, lognormal, and normal distribution testing
2008	Sohn and Kim	Seoul, Korea	N/A	Beacon-based probing system	N/A	Travel time index Buffer index	Autoregressive moving average-generalized autoregressive conditional heteroscedasticity (ARMA-GARCH) model
2010	Saberi and Bertini	Portland, Oregon, US	5-min	Inductive loop detectors	3 p.m. – 6 p.m.	Buffer index, Coefficient of variation	Standard midpoint algorithm
2011	Hainen et al.	Indiana, US	N/A	Bluetooth-enabled devices	N/A	25th, 50th, and 75th percentile travel time	N/A
2012	Yazici et al.	New York City, US	N/A	NYC Taxis GPS data	N/A	Coefficient of variation, λ^{skew}, λ^{width}	Commission and employed classification and regression tree (C&RT) methodology
2016	Zheng et al.	Changsha, China	30s	ANPR data	0:00 a.m.- 23:59 p.m.	Standard deviation, travel time skewness.	Zuylen’s delay probability distribution model

2.3.2. Network Level Travel Time Variability Analysis

Research studies that model network level travel time variability using TTR measures are reviewed in this section. Table 2.4 provides a summary of the studies reviewed in this section in chronological order.

2.3.2.1. Yang et al.’s research work

Yang et al. (2014) utilized the Hasofer–Lind–Rackwitz–Fiessler algorithm which was widely used in the field of reliability engineering to calculate the reliability index of a system. The modeling framework consisted of three parts: travel time estimation, travel time distribution estimation, and corridor-network TTR index calculation. A description of the data set used in this study was followed by the implementation and applications of the proposed method. The results showed that this modeling method could better capture the variability of traffic flow in detail, especially during rush hours.

2.3.2.2. Recker et al.’s research work

Recker et al. (2005) conducted a study on risk-taking route choice via the analyses of travel time variability data of section, corridor, and network under different demand levels. The TTR was also evaluated. In this study, path TTR was defined as “*the probability that the travel time of a given path is within an acceptable threshold.*” OD-TTR was defined as “the *probability that the weighted average travel time of a given OD pair is within an acceptable threshold.*” The evaluation procedure was based upon a Monte Carlo simulation framework. Three scenarios were constructed to test how different route choice models affect the estimation of travel time reliability under uncertain environment. The analysis can be concluded as: “*as the degree of risk aversion to network uncertainty increases, travel time also increases and results in lower travel time reliability*.”

2.3.2.3. Clark and Watling's research work

Clark and Watling (2005) conducted a study to estimate the total network travel time probability distribution. They considered day-to-day demand variations in the travel demand matrix as a main factor affecting travel time variability and estimated the total travel time density function. The numerical test results indicated that the application of this approach was suitable to understand the impact of capacity changes.

2.3.2.4. Ng and Waller's research work

Ng and Waller (2010) developed a methodology to assess TTR in a transportation network under uncertain road capacities. A Fourier transformation method was presented in this study. The special case when capacities were normally distributed random variables was also considered. The proposed method was applied to analyze the impact of capacity variations on the TTR, which was proved to be valid.

2.3.2.5. Tu et al.'s research work

Tu et al. (2013) investigated a macroscopic TTR diagram to relate the TTR to the network density. The authors conducted empirical analyses to investigate the variability in macroscopic fundamental diagram (MFD) as seen in scatter plots using traffic data of freeway systems in Netherlands. A critical TTR accumulation point was found to exist, *"below which network accumulation had little impact on travel time reliability and had a significant impact when it is above"*. The critical TTR accumulation was also found to be usually lower than the critical MFD accumulation.

TABLE 2.4: Summary of network level travel time variability studies

Year	**Author**	**Location**	**Data Aggregation**	**Data Source**	**Study Periods**	**TTR Measure(s)**	**Modeling Algorithm**
2004	Recker et al.	Orange County, US	5-min	Loop detector data	4 - 10 a.m.	Standard deviation	Mixed logit route choice model
2005	Clark and Watling	N/A	N/A	N/A	N/A	100(1 – Pr(M > 5))%	Poisson demand distribution model; Monte Carlo method
2010	Ng and Waller	N/A	N/A	N/A	N/A	Volume-to-capacity ratios	Fourier transforms, bureau of public roads (BPR) function;
2013	Tu et al.	Netherlands	10-min	dual loop detectors	N/A	Probability of traffic breakdown	Piecewise linear speed-based trajectory algorithm;
2014	Yang et al.	St. Louis, US	15-min	MODOT traffic sensors' data	9 a.m.- 5 p.m.	Standard deviation	Kernel density estimation technique

2.3.3. Travel Time Variability Analysis with the Consideration of Incidents and Weather

Research studies on travel time variability with the consideration of incidents and weather were reviewed and summarized in this section. Table 2.5 provides a summary of the studies reviewed in this section in chronological order.

2.3.3.1. Hojati et al.’s research work

Hojati et al. (2016) developed a method to quantify the impact of traffic incidents on TTR on freeways. The authors first obtained the recurrent speed profile for each specific link and DOW using the Quantum-Frequency Algorithm. The non-recurrent congestion was identified as an ‘event’ with a start time and end time. Next, the total travel time due to an event on a set of affected links was modeled, and then the BI was selected as the TTR measure. The authors then conducted a Tobit regression analysis which can handle the presence of censored data either in the lower tail or in the upper tail. Based on the Queensland DOT and STREAMS Incident Management System (SIMS) database, 430 incidents were matched with the identified events. Finally, 3 Tobit model estimation results were shown focusing on crashes, hazards and stationary vehicles.

2.3.3.2. Charlotte et al.’s research work

Charlotte et al. (2017) presented an empirical analysis of travel time distribution on urban roads in the region of Paris, France. Historical data of accidents and roadway works were added to evaluate the impact of some non-recurrent influencing factors. 90th percentile of the travel time distribution was modeled with linear models including explanatory variables including number of lanes, mean value of the travel time distribution, travel direction, time of the day, number of accidents and roadworks.

2.3.3.3. Martchouk et al.’s research work

Martchouk et al. (2010) studied the travel-time variability with the travel-time data on freeway segments in Indianapolis collected with the help of anonymous Bluetooth sampling techniques. The effects of adverse weather were discussed in the study. The results showed that the travel time increased during adverse weather period, and the variance in travel times during the same time period also increased. Various statistical models were also estimated in the study to understand the effect of individual vehicle travel times variability as well as average travel times variability. For the individual vehicle travel time model, the probability of travel duration time changes of a segment was estimated. As anticipated, higher average speed led to lower individual travel time, whereas higher distance and volume resulted in increased travel time. In the average travel time model, estimated parameter indicated that higher average travel time during the previous time period resulted in higher average travel time during the current period.

2.3.3.4. Peer et al.’s research work

Peer et al. (2012) conducted a study to provide simple rules to predict travel time variability based on the travel time data of 145 (one-directional) highway links in Netherlands. Standard deviation of travel times was used as the TTR measure. The explanatory variables included DOW, season, weather condition and network condition. Formulas for TTR were built based on ‘rough information’ and ‘fine information’. Mean delay was also analyzed to express the travel time.

The empirical analysis of travel time variability results showed that *a shorter link is on average associated with lower variability*. The authors also found that variability is positively correlated with the number of lanes for smaller delays and it is negatively correlated with the number of lanes for longer delays.

2.3.3.5. Shao et al.'s research work

Shao et al. (2008) proposed a new travel time reliability-based stochastic user equilibrium traffic assignment model to investigate the effects of rain on risk-taking behaviors of different road users in networks with day-to-day demand fluctuations and variations in travel time. To capture the rain effects on travel time, a new travel time function was developed based on the conventional BPR function. Rain effects on traffic demand were also modeled via the conventional elastic demand function. Finally, it was found in the numerical results that path choice behaviors and traffic demand of different road users were affected by the rainfall intensity.

2.3.3.6. Li et al.'s research work

Li et al. (2016) conducted a study which focused on studying the weather impact on traffic operations. Different rainfall intensity data for every hour of Florida regions were incorporated into the TTR model along with the historical speed database. Different scenarios for each hour (under clear weather, light rain, and heavy rain conditions) were created and applied to the respective roadway sections. The results showed that the speed reductions on arterials were 10% for light rain and 12% for heavy rain. However, the assumed reduction in the speed on arterials caused by rain intensity may need to be verified with additional empirical data during a long period of time to reveal the trends and impacts with more confidence and accuracy.

2.3.3.7. Kamga and Yazici's research work

Kamga and Yazici (2014) conducted a study via merging taxi trips' GPS records and historical weather records of New York City and then calculated the descriptive statistics of travel time for different TOD, DOW and various weather conditions. The weather conditions were categorized into 8 groups including Clear, Light rain, Rain, Heavy rain, Light snow, Snow, Heavy Snow and Unknown. Based on the value of each coefficient, the Classification and Regression Trees (C&RT) model was used to extract the travel time coefficients distribution under each DOW-TOD-Weather category.

The temporal pattern analysis results of each travel time parameter were finally presented. With the analysis results of CV, the authors pointed out: "*Regarding the weather impacts, it was found that inclement weather indeed increases average travel times yet decreases variability, resulting in higher travel reliability indicated by lower coefficients of variation*."

TABLE 2.5: Summary of travel time variability studies with the consideration of weather/incident impacts

Year	Author	Location	Data Aggregation	Data Source	Study Periods	TTR Measure(s)	Modeling Algorithm
2008	Shao et al.	N/A	N/A	N/A	N/A	Standard deviation	Reliability-based stochastic user equilibrium
2011	Martchouk et al.	Indiana, US	N/A	Microwave detectors	N/A	Standard deviation	Hazards-based model BPR function
2012	Peer et al.	Netherlands	15-min	Loop detector	6:00 a.m. - 8:15 p.m.	Standard deviation	Non-traffic regime-based model; traffic regime-based model
2014	Kamga and Yazici	NYC, US	N/A	Taxis GPS data	N/A	Standard Deviation, Average travel time, Coefficient of variation	Classification and Regression Trees (C&RT) model
2016	Li et al.	Florida	N/A	FDOT database	N/A	Actual travel time	FDOT travel time reliability model
2016	Hojati et al.	Queensland	N/A	Queensland traffic department	N/A	Extra buffer time index	Tobit model
2017	Charlotte et al.	Paris	6-min	Loop detectors	7 - 10 a.m. 5 - 8 p.m. 10p.m. - 5 a.m.	90th percentile travel time	N/A

2.3.4. Travel Time Variability Analysis with the Consideration of Multiple Influencing Factors

Research studies on travel time variability with the consideration of multiple influencing factors were reviewed and summarized in this section. Table 2.6 provides a summary of the studies reviewed in this section in chronological order.

2.3.4.1. Tu's research work

Tu (2008) developed a TTR model with the consideration of four influencing factors including road geometry, adverse weather, speed limits, and traffic accidents. The model was validated using traffic data from urban freeways in Netherlands. The results of road geometry impacts indicated that there was a threshold value L for the length of ramp/weaving section. If the actual length was less than L, the TTR would decrease with the decreasing length of ramp/weaving sections. If the actual length was larger than L, the length has far less impact on travel time reliability. TTR on the freeway was also strongly impacted by the number of ramps per unit road length. Above a threshold value, the more ramps contribute to the lower TTR. The results of adverse weather's impacts indicated that adverse weather conditions clearly have negative effects on TTR on the freeway, which means that travel times are less reliable under adverse weather conditions than those under normal weather conditions, especially at higher inflow levels.

2.3.4.2. Javid and Javid's research work

Javid and Javid (2017) developed a framework to estimate travel time variability caused by traffic incidents based on integrated traffic, road geometry, incident, and weather data. A series of robust regression models were developed based on the data from a stretch in California's highway system. Next, travel time variability was estimated via the proposed speed change models, and the results were compared with the actual changes in travel time. The results of the split-sample validation showed the effectiveness of the proposed models in estimating the travel time variability. In conclusion, for incidents occurring on weekends, the highway clearance time would be shorter. Shoulder existence and lane width would adversely impact downstream highway clearance time.

2.3.4.3. Schroeder et al.'s research work

Schroeder et al. (2013) presented a methodology for freeway travel time analysis based on freeway data in North Carolina. The variability impact considerations included time-of-day, day-of-week, and month-of-year differences, and various nonrecurring congestion sources (such as weather, incidents, work zones, and special events). The freeway scenario generator was used and resulted in 2,508 scenarios based on freeway facility data in North Carolina. The resulting travel time distribution was presented, and a sensitivity analysis was conducted to explore the relationship between weather and incidents and the overall reliability of the facility.

2.3.4.4. Kwon et al.'s research work

Kwon et al. (2017) developed an empirical corridor level method to study the travel time variability. The authors divided the variables which had an impact on the travel time into three categories: traffic influencing events (traffic incidents and crashes, work zone activity, weather and environmental conditions), traffic demand (fluctuations in day-to-day demand and special events), and physical road features (traffic control devices and inadequate base capacity). A linear regression statistical model was then constructed to conduct the travel time reliability analysis. Buffer time (95th percentile of travel time - median travel time) was chosen over other measures to represent the TTR because it was more popular and easier to formulate and fit the model. The model was tested in San Francisco Bay Area and used to identify how each variable contributes to the TTR. The results of this study provided useful insights into predicting the TTR.

2.3.4.5. Kim's research work

Kim (2014) conducted a study on freeway travel time variability and developed a compound Gamma distribution model. The model captured both vehicle-to-vehicle and day-to-day travel delay. The author also proposed a framework that features scenario-based simulation approaches. Factors such as incidents, bad weather, work-zone, and planned special events were considered in this study. This approach could provide the ability to forecast potential variations in travel time and estimate travel time distributions with more accuracy.

TABLE 2.6: Summary of travel time variability studies with the consideration of multiple influencing factors

Year	Author	Location	Data Aggregation	Data Source	Study Periods	TTR Measure(s)	Modeling Algorithm
2008	Tu	Delft, Netherlands	N/A	Regiolab-Delft monitoring system.	N/A	λ^{skew}, λ^{width}	Dynamic traffic assignment
2013	Schroeder et al.	Durham, NC, US	15-min	INRIX	2-8 p.m.	Travel time index	Highway capacity manual model
2014	Kim	NYC, US	5-min	ASOS station, INFORM system	6-10 a.m.	Percent variation, Misery index and Buffer time index	Expectation-maximization algorithm
2017	Kwon et al.	San Francisco, CA, US	N/A	Performance measurement system (PeMS)	7 - 9 a.m. 11 a.m. - 1 p.m. 4-6 p.m.	Buffer time	Linear regression
2017	Javid and Javid	California, US	5-min	PeMS database	N/A	Index of agreement, Correlation coefficient	Robust regression

2.4. Travel Time Prediction Using Machine Learning Approaches

Traditionally, the methods such as linear regression and time series models have been widely applied to predict travel times using historical data. However, with the consideration of effectiveness, accuracy and feasibility, these models may become outdated and replaceable. With the development of artificial intelligence technologies, various novel prediction methods have been developed accordingly in recent years. With the help of ITS systems and the traffic data, different machine learning approaches have been deployed in the travel time prediction area. The methodology can include, but are not limited to: Support vector machine regression, Neural network approaches (e.g. State-and-space neural network, long short term memory neural network), nearest neighbor (e.g. k-nearest neighbor), and ensemble learning (e.g. Random Forest and gradient boosting), etc. The review of different approaches will be helpful to find the most appropriate, advanced and accurate model in this study. Research studies that used machine learning/deep learning methods to predict travel time are reviewed and summarized in this section. Table 2.7 provides a summary of the studies reviewed in this section in chronological order.

2.4.1. Support Vector Regression Approach

2.4.1.1. Wu et al.'s research work

Wu et al. (2004) applied SVR for travel-time prediction and compared its results to other baseline travel time prediction methods using real world highway traffic data. Since support vector machines have greater generalization ability and can guarantee global minima for given training data, it was believed that SVR would perform well for time series analysis. The results showed that the SVR model can "*significantly reduce both relative mean errors and root-mean-squared errors of predicted travel times*". This study demonstrated the feasibility of applying SVR in travel time prediction and proved that SVR is applicable for traffic data analysis.

2.4.2. Neural Network Approach

2.4.2.1. Park and Rilett's research work

Park and Rilett (1999) proposed a BP neural network model to predict freeway link travel time. The freeway link travel time data collected on the freeway of Houston, Texas, by the automatic vehicle identification (AVI) system were used as the validation database. The proposed model can provide acceptable prediction results with the mean absolute percentage error (MAPE) being ranged from 7.4% to 18%.

2.4.2.2. Van Lint et al.'s research work

Van lint et al. (2002) presented an approach to predicting freeway travel time based on the state-space neural network. The data from freeway operations simulation (FOSIM) 4.1 were used to train and test the travel time prediction model. The authors also eliminated the insignificant parameters in the model and made it more effective without the loss of predictive performance.

2.4.2.3. Wisitpongphan et al.'s research work

Wisitpongphan et al. (2012) proposed a back propagation (BP) neural network model to predict freeway link travel time. The one-month vehicle trajectory data of 297 probe vehicles via GPS database in Thailand were used as the validation database. The prediction results of the proposed model can accurately approximate the travel time with the mean squared error (MSE) being less than 3%.

2.4.2.4. Zheng and Van Zuylen’s research work

Zheng and Van Zuylen (2013) conducted a study using the probe vehicle data to estimate complete link travel times. Based on the information collected by probe vehicles, a three-layer neural network model was developed by the authors to estimate complete link travel time for individual probe vehicle traversing the link. The estimation result of this model was then compared with that of an analytical estimation model. The performance of these two models were evaluated using the data derived from VISSIM simulation model. The final results suggested that the Artificial Neural Network model performs better.

2.4.2.5. Duan et al.’s research work

Duan et al. (2016) employed a long short-term memory (LSTM) neural network model to predict freeway travel time. The authors constructed 66 series LSTM neural networks by using travel time data collected along 66 links of the highways in England. The authors discussed the predictions of multi-step ahead travel time and found 1-step ahead travel time prediction can provide best results.

2.4.2.6. Liu et al.’s research work

Liu et al. (2017) proposed a LSTM deep neural network model using 16 settings of hyper-parameters to predict the travel time on the interstate highways in California, U.S. The results of proposed model were compared with the results of other regression models and Autoregressive integrated moving average (ARIMA) model and showed that the performance of the LSTM neural network model was the best.

2.4.2.7. Wang et al.’s research work

Wang et al. (2018) presented a machine learning-based method to predict the vehicle travel time using floating-car data. The authors adapted different machine learning models to solve the regression problem. Furthermore, the authors evaluated the solution offline with millions of historical vehicle travel data and the results showed that their proposed deep learning algorithm significantly outperforms the other state-of-the-art algorithms.

2.4.2.8. Wang et al.'s research work

Wang et al. (2018) proposed a LSTM neural network-based travel time prediction model using the historical vehicle trajectory data. Both road segment-based travel time estimation and path-based travel time estimation were discussed in this study. The results showed that the proposed model can effectively capture the spatial and temporal dependencies and accurately predict travel time.

2.4.2.9. Wei et al.'s research work

Wei et al. (2018) combined the convolutional neural network and LSTM neural network together to predict the short-term travel time. The vehicle trajectory data on the urban roads were used in this study. The author pointed out that the prediction of the proposed model was more effective than that of other existing approaches.

2.4.3. Nearest Neighbors Approach

2.4.3.1. Yu et al.'s research work

Yu et al. (2017) combined the Random Forest model and K-NN model in their study to predict bus travel time. The proposed combined-model was compared with linear regression, K-NN, SVM and Random Forest. The results showed the proposed model achieved highest accuracy level and can be applied to real-time prediction.

2.4.3.2. Myung et al.’s research work

Myung et al. (2011) proposed a model to predict travel times with the help of k nearest neighbor (KNN) method using data obtained from vehicle detector system and the automatic toll collection system. The model combined these two datasets and minimized the limitations of each dataset. The authors compared the prediction results of the proposed model with other models using actual travel time data. The comparison results showed that the proposed model performs much better than other models.

2.4.3.3. Moonam et al.’s research work

Moonam et al. (2019) conducted a study to predict freeway travel time based on the experienced travel time using several methodologies including k-nearest neighbor (k-NN), least squares regression boosting and Kalman filter (KF) methods. The authors compared the performances of each methods from both link and corridor perspectives and pointed that “*the KF method offers superior prediction accuracy in a link-based model*”.

2.4.4. Ensemble Learning Approach

2.4.4.1. Hamner et al.’s research work

Hamner et al. (2011) applied a context-dependent Random Forest method to predict travel-time based on GPS data of the cars on the road in a simulation framework. The root mean squared error (RMSE) of the model was less than 7.5%.

2.4.4.2. Zhang and Haghani’s research work

Zhang and Haghani (2015) employed a gradient boosting regression tree method to analyze and predict freeway travel time to improve the prediction accuracy. The authors used travel time data along freeway sections in Maryland and discussed the effects of different parameters on the proposed model and the correlations of input and output variables. The prediction results showed the proposed model can provide considerable advantages in freeway travel time prediction.

2.4.4.3. Li and Bai's research work

Li and Bai (2016) employed a gradient boosting regression tree method to analyze and predict travel time of freight vehicles. The authors used travel time data and vehicle trajectory data in Ningbo, China. The prediction results showed the proposed model can be feasible in the real-world.

2.4.4.4. Fan et al.'s research work

Fan et al. (2017) conducted a study using the Random Forest method to predict freeway travel time with the help of data collected from highway electronic toll collection in Taiwan. The results can help drivers to select optimal departure times to avoid traffic congestion and thus minimize travel time.

2.4.4.5. Gupta et al.'s research work

Gupta et al. (2018) employed Random Forest and gradient boosting models to predict taxi travel time in Porto, Portugal. The vehicle trajectory data were used as the database and it was found that the gradient boosting model provided better prediction results than the Random Forest model.

TABLE 2.7: Summary of travel time prediction using machine learning approaches

Year	Author	Location	Roadway Category	Data Source	Data Type	Prediction method
1999	Park and Rilett	Houston, US	Highway	AVI system	Travel time	BP Neural Network
2002	Van Lint et al.	N/A	Freeway	FOSIM (freeway operations simulation)	Travel time, travel speed	State-Space Neural Network
2005	Wu et al.	Taiwan	Highway	Loop detector	Travel speed	SVR
2010	Hamner et al.	N/A	N/A	Global Positioning System (GPS)	Travel speed	Random Forest
2011	Myung et al.	Korea	N/A	Automatic traffic count system	Travel time	KNN
2012	Wisitpongphan	Bangkok, Thailand	Highway	GPS	Travel time, GPS	BP Neural Network
2013	Zheng and Van Zuylen	Delft, Netherlands	Urban road	GPS data	Vehicle position, travel speed	State-Space Neural Network
2015	Zhang and Haghani	Maryland, US	Interstate highway	INRIX Company	Travel time	Gradient boosting
2016	Duan et al.	England	Highway	Cameras, GPS and loop detectors	Travel time	LSTM Neural Network
2016	Li and Bai	Ningbo, China	N/A	N/A	Truck trajectory, travel time, travel speed	Gradient boosting
2017	Liu et al.	California, US	Interstate highway	PeMS	Travel time	LSTM Neural Network
2017	Fan et al.	Taiwan	Highway	Electric toll	Travel time, vehicle information	Random Forest
2017	Yu et al.	Shenyang, China	Bus route	Automatic Vehicle Location system	Bus travel time	Random Forest and K-NN
2018	Wang et al.	Beijing, China	Urban road	Floating Car Data	Taxi travel time, vehicle trajectory data	LSTM Neural Network
2018	Wei et al.	China	Urban road	Vehicle passage records	Travel time	LSTM Neural Network
2018	Wang et al.	Beijing and Chengdu, China	Urban road	GPS	Vehicle trajectory data	LSTM Neural Network
2018	Gupta et al.	Porto, Portugal	Urban road	GPS	Taxi travel speed	Random forest and gradient boosting
2019	Moonam et al.	Madison, Wisconsin, US	Freeway	Bluetooth detector	Travel speed	K-NN, KF

2.5. Summary

A comprehensive review of the current and historical researches related to TTR definitions, measures, travel time variability analysis and machine learning-based travel time prediction methodologies has been presented in the preceding sections. This is intended to provide a solid reference and assistance in analyzing travel time variability and developing travel time prediction models.

CHAPTER 3: DATA DESCRIPTION AND PROCESSING

3.1. Introduction

This chapter provides the basic information needed to analyze travel time variability and conduct travel time prediction, including the travel time data and historical weather data utilized in this study. The following sections are organized as follows. Section 3.2 presents detailed information about the raw travel time data source, followed by the discussions about weather data collection in section 3.3. Section 3.4 described details of data processing. Finally, section 3.5 concludes this chapter with a summary.

3.2. Travel Time Data Collection

This study focuses on the travel time data gathered from the Regional Integrated Transportation Information System (RITIS) website and uses the collected data to conduct the TTR analysis and travel time prediction. A series of major freeway segments are selected for the case study: Interstate 77 (I-77) Southbound (Figure 3.1) is one of the most heavily traveled Interstate highways in Charlotte, North Carolina and runs from north to south. All the selected segments have uninterrupted coverage of RITIS data 24 hours per day and 365 days a year.

Interstate 77 begins at the South Carolina state line, near Fort Mill, and goes through the city of Charlotte as a major north-south corridor, connecting the Charlotte center area with the suburbs of Pineville, Huntersville, Cornelius, and Davidson. The highways in Charlotte area experience massive traffic congestion during weekdays due to heavy commuter and interstate traffic.

The selected section of I-77 Southbound starts from the intersection with Harris oak Blvd and ends at the interchange with I-485 (Exit 2) at the south part of the city. 32

roadway segments are selected in this study, and the total length of the selected section is 19 miles.

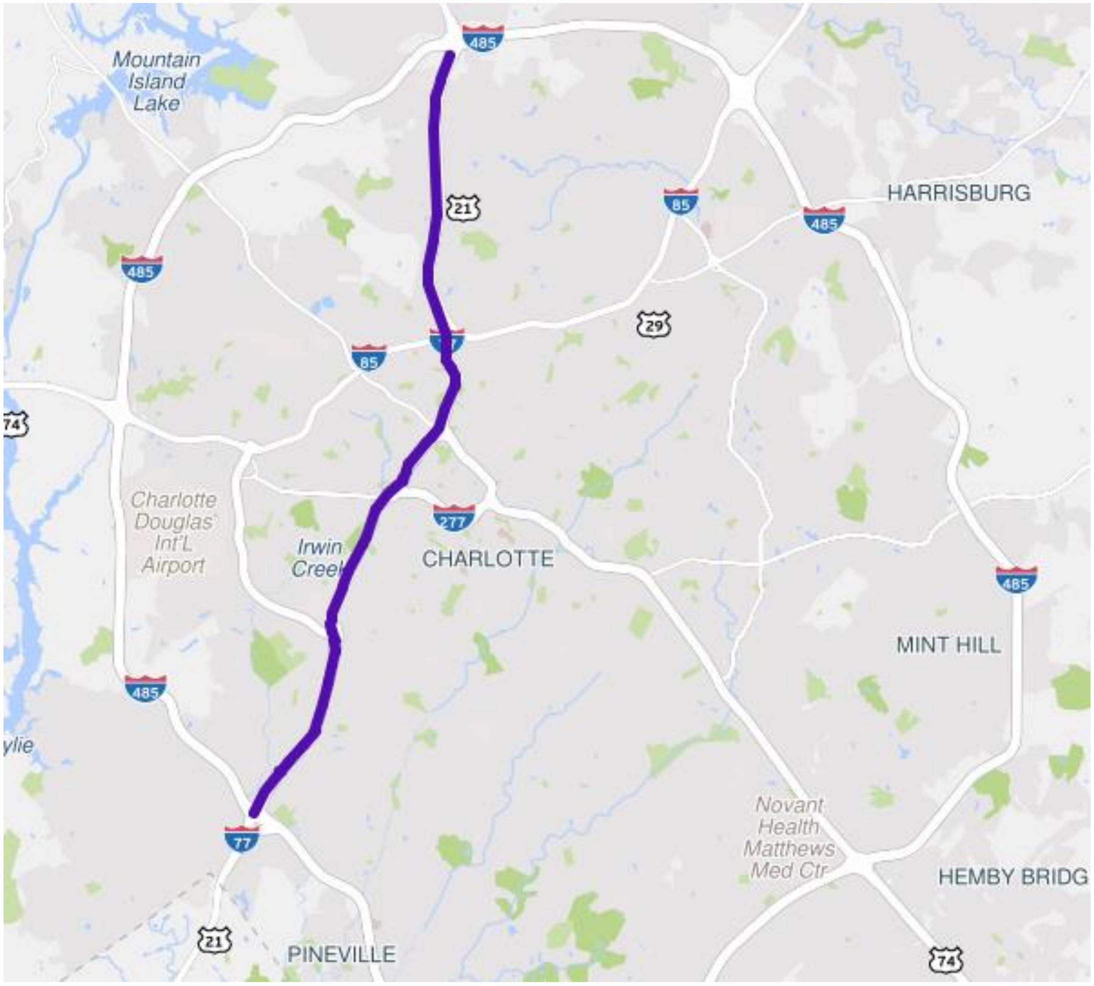


FIGURE 3.1: Selected I-77 southbound section

As discussed in the literature review, in the past, travel time was deduced from the loop detector data, historical trends or floating car runs. In this study, travel time and speed data are obtained from the RITIS website which gathered information about roadway speeds and vehicle counts from 300 million real-time anonymous mobile phones, connected cars, trucks, delivery vans, and other fleet vehicles equipped with GPS locator devices.

On the RITIS website probe data analytic suite, the raw probe data can be downloaded with the desired section and format. The roadway section can be selected based on the Road states and countries, Traffic message channels (TMCs), Directions, Zip codes, Road class and Road name. The partial sections can be selected with the selection of begin and end intersections. The date range can be selected from January 1st, 2008 to today. Seven days of week and times of day from 12:00 AM to 11:59 PM can also be selected. The units of travel time can be categorized into both seconds and minutes. The averaging period can be selected as five minutes, ten minutes, fifteen minutes and one hour. A sample of raw travel time data utilized in this study is shown in Table 3.1 below:

TABLE 3.1: Sample raw travel time data

TMC Code	Measurement_tstamp	Speed	Travel_time_seconds
125N04784	1/1/2015 0:00	62.91	53.58
125N04783	1/1/2015 0:00	61.17	12.82
125N04786	1/1/2015 0:00	60.43	47.56
125N04785	1/1/2015 0:00	61.30	11.85
125N04780	1/1/2015 0:00	63.97	14.59
125N04782	1/1/2015 0:00	63.04	21.73
125N04781	1/1/2015 0:00	62.79	12.42
125N04788	1/1/2015 0:00	65.03	29.60
125N04787	1/1/2015 0:00	63.50	53.76
125N04789	1/1/2015 0:00	64.79	54.50
125-04783	1/1/2015 0:00	62.98	33.22
125-04782	1/1/2015 0:00	62.75	35.68
125-04785	1/1/2015 0:00	60.54	5.16
125N04784	1/1/2015 0:00	62.91	53.58

Table 3.1 contains the following information:

TMC_Code: The RITIS Probe Data Analytics Suite uses the TMC standard to uniquely identify each road segment. This field indicates the segment ID.

Measurement_tstamp: This field indicates the timestamp of the record.

Speed: This field indicates the current estimated harmonic mean speed for the roadway segment in miles per hour.

Travel_time_seconds: This field indicates the time it will take to drive along the roadway segment.

3.3. Weather Data Collection

The historical weather data near the Charlotte Douglas International airport can be found at the www.wunderground.com website. The raw weather data can be achieved within the desired time period. The date range can be selected from January 1st, 1941 to today.

The raw weather data include information on different categories such as temperature, dew point, humidity, pressure, visibility, wind direction, wind speed, gust speed, precipitation, and conditions. The raw weather data from this website were recorded per hour. Due to the discrepancy in the time interval, one-to-one mapping or correlation study cannot be done using the original data. Hence, the methodology to combine the traffic data with the weather data will be discussed in the next section. The sample of weather data achieved is shown in Table 3.2 below.

TABLE 3.2: Sample raw weather data

Date	Time (EDT)	Visibility	Conditions
Saturday, March 14, 2009	6:55 AM	2.0 miles	Rain
Saturday, March 14, 2009	7:55 AM	2.0 miles	Rain
Saturday, March 14, 2009	8:55 AM	2.0 miles	Light Rain
Saturday, March 14, 2009	9:55 AM	2.0 miles	Light Rain
Saturday, March 14, 2009	10:55 AM	3.0 miles	Light Rain
Saturday, March 14, 2009	11:55 AM	2.0 miles	Light Rain
Saturday, March 14, 2009	12:55 PM	3.0 miles	Light Rain
Saturday, March 14, 2009	1:55 PM	7.0 miles	Light Rain
Saturday, March 14, 2009	2:55 PM	6.0 miles	Light Rain
Saturday, March 14, 2009	3:55 PM	7.0 miles	Light Rain
Saturday, March 14, 2009	4:55 PM	4.0 miles	Rain

3.4. Data Processing

Based on previous studies, it is widely accepted that only severe weather events will cause a significant impact on speeds and travel times. Due to the weather characteristics in the Charlotte area and the distribution of each weather category, detailed weather conditions are categorized into three groups including normal, rain, and snow/fog/ice. Table 3.3 presents the detailed classification of the weather conditions. Conditions such as “overcast” or “mostly cloudy” are assumed to be no different from “clear” conditions due to no obvious impact on traffic conditions. These conditions are categorized into ‘normal’. All the conditions such as ‘rain’ or ‘thunderstorm’ are categorized as ‘rain’. In order to ensure the acceptable sample size, “snow”, “fog”, “ice pellet”, and other similar conditions are combined together due to their rate of occurrence.

TABLE 3.3: Classification of the weather conditions

Original Weather Condition	New Weather Category
Haze	Snow/fog/ice
Fog	
Smoke	
Patches of Fog	
Mist	
Shallow Fog	
Light Freezing R	
Light Ice Pellet	
Light Freezing D	
Light Freezing F	
Ice Pellets	
Light Snow	
Snow	
Heavy Snow	
Clear	Normal
Partly Cloudy	
Mostly Cloudy	
Scattered Clouds	
Overcast	
Unknown	
Light Rain	Rain
Rain	
Heavy Rain	
Light Drizzle	
Heavy Thunderstorm	
Light Thunderstorm	
Thunderstorm	
Drizzle	
Squalls	

Figure 3.2 illustrates the data processing steps. In order to merge the link travel times dataset with historical weather dataset, the issue of different intervals of two datasets should be resolved first. The RITIS datasets are aggregated into 15-minute intervals, while the weather dataset is aggregated into one-hour intervals. Therefore, the weather conditions are distributed evenly with RITIS dataset based on the timestamp.

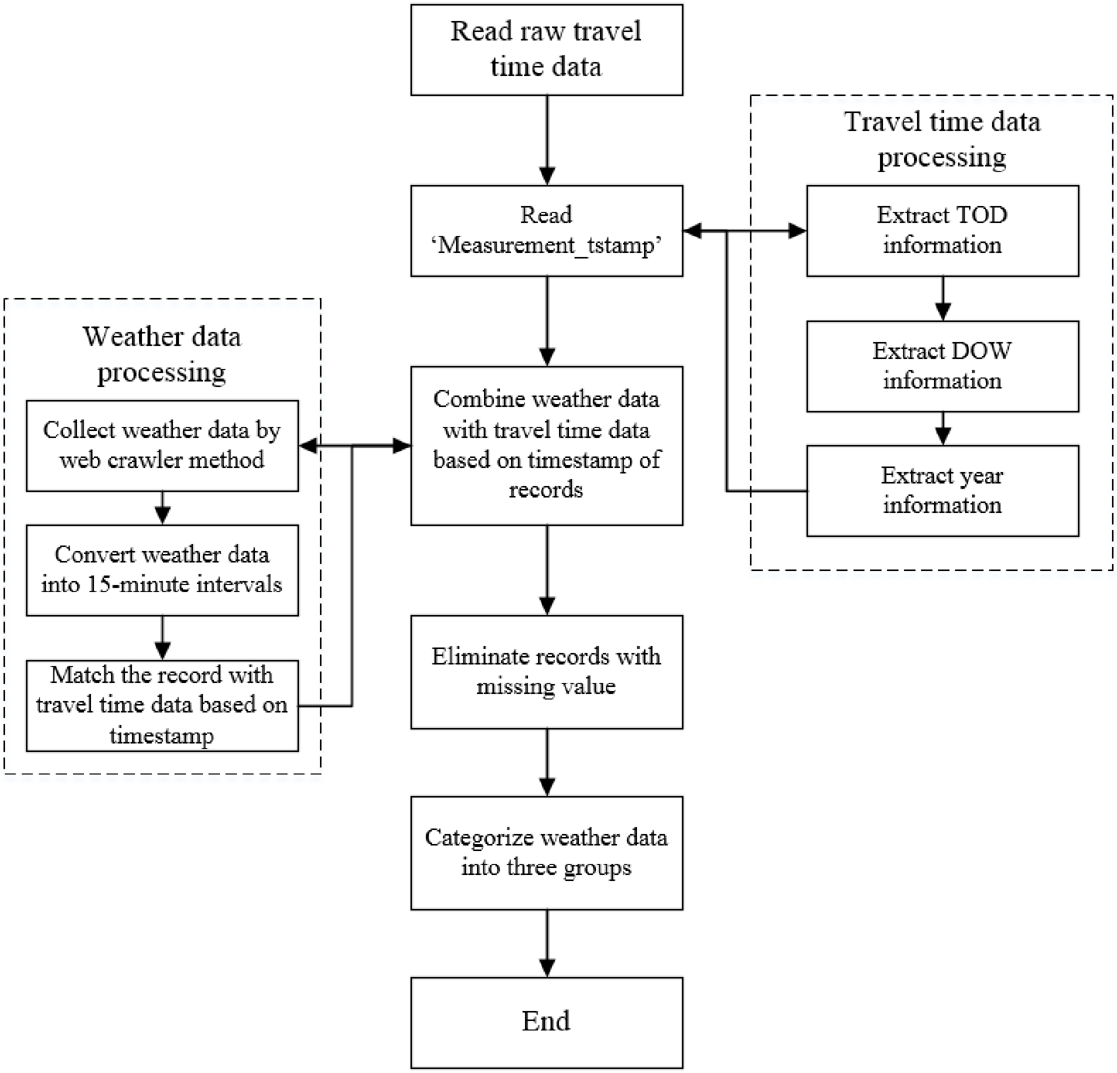

Read raw travel time data
Read 'Measurement_tstamp'
Travel time data processing
Extract TOD information
Extract DOW information
Extract year information
Weather data processing
Collect weather data by web crawler method
Convert weather data into 15-minute intervals
Match the record with travel time data based on timestamp
Combine weather data with travel time data based on timestamp of records
Eliminate records with missing value
Categorize weather data into three groups
End

FIGURE 3.2: Data processing flow chart

3.5. Summary

This chapter presents the detailed information on the data source, data structure, and processing methodology to combine the travel time with raw weather data. This is intended to provide a solid reference and assistance in analyzing travel time variability for future tasks.

CHAPTER 4: TRAVEL TIME VARIABILITY ANALYSIS

4.1. Introduction

The chapter presents the analysis of travel time variability patterns. The following sections are organized as follows. Section 4.2 shows the study location identification process based on the TTR. Section 4.3 presents the travel time variability patterns under all conditions. Section 4.4 discusses the travel time variability patterns considering the DOW. Section 4.5 describes the travel time variability patterns considering different weather conditions. Finally, section 4.6 concludes this chapter with a summary.

4.2. Study Location Identification Based on TTR

This section describes how to identify study locations based on the TTR measure. The indicator is calculated by aggregating the speed and travel time observations collected during the time interval of interest across a year. A number of performance measures such as FOC, PTI, BI can be applied to achieve this goal. For illustration purpose and other reasons that will be discussed later, we only present how to extract the PTI values for each segment during each time interval.

4.2.1. Selection of TTR Measures

TTR measures have been increasingly encouraged by FHWA for use to manage and operate transportation systems. Previous research has led to the employment of various TTR measures to assist in highway performance evaluation and congestion management. In the literature review chapter, we have introduced different types of travel time reliability measures such as the 95th percentile travel time, BI, PTI, MI, CV, FOC, skew of travel time distribution, and width of travel time distribution.

There are four most widely used TTR measures in previous studies and they are BI, PTI, CV, and FOC. However, BTI and CV have the limitation since their values depend on the average travel time, which may change over time (Fan and Gong, 2017). Therefore, the PTI is chosen as the primary measure of travel time reliability in this study. It is calculated by dividing 95th percentile travel time by the free flow travel time so as to represent the percentage of extra travel time that most people will need to add on to their trip in order to ensure on-time arrival. For example, a PTI value of 1.5 at 5 PM means that for a 20-minute trip in light traffic, 30 minutes should be planned at 5 PM to make sure that he or she is on time. The equation of PTI is provided below:

$$PTI_i = \frac{T_{i95}}{FFTT_i}$$

where,

PTI_i = The planning time index of segment *i*.

T_{i95} = 95th percentile travel time on the TMC segment *i* during the study period across multiple days (e.g., a month) or a year.

$FFTT_i$ = Free-flow travel time on TMC *i* during the same observation period as mentioned above.

For each roadway segment, the free-flow travel time is computed by dividing the length of segment by the free-flow speed, which was defined as the 85th percentile speed during overnight hours (10 p.m. to 5 a.m.) (Florida DOT, 2011, Schrank et al. 2015, Fan and Gong, 2017).

4.2.2. Corridor PTI Information Aggregation

The first step to identify the study segments is to plot the two-dimensional PTI matrix for each road segment along the corridor. This would provide a straightforward and visualized tool for decision-makers to grasp the average traffic conditions along a corridor. The long-term (in one-year period) PTI values of each segment from 2011 to 2015 were calculated and shown in Figure 4.1 to Figure 4.5, respectively. Note that in these figures, the horizontal axis denotes the time of day and the vertical axis represents TMC segments along the selected section on I-77 Southbound. Each cell represents the PTI value. The darker the color, the higher the PTI.

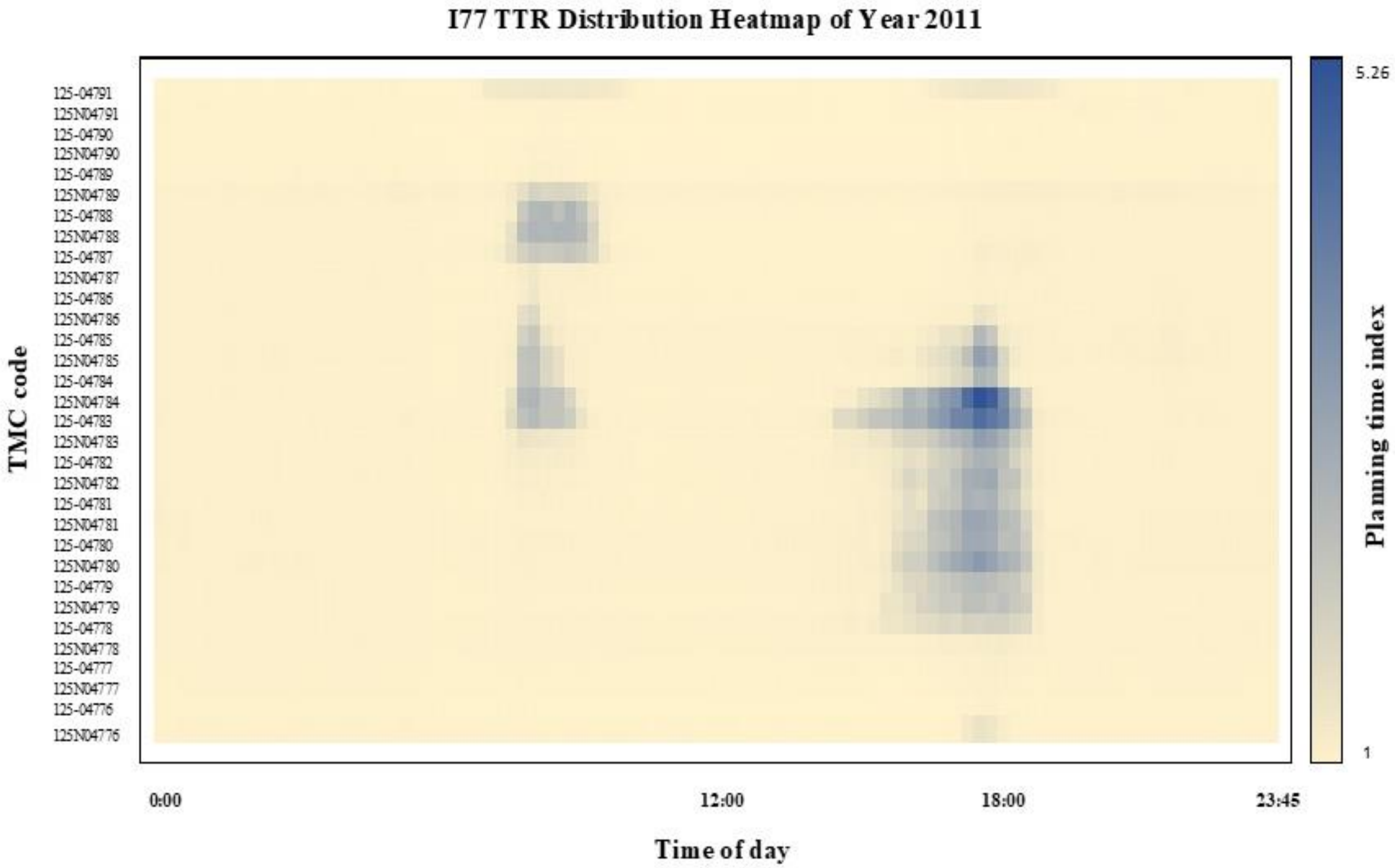


FIGURE 4.1: PTI heatmap of I77 (SB) in year 2011

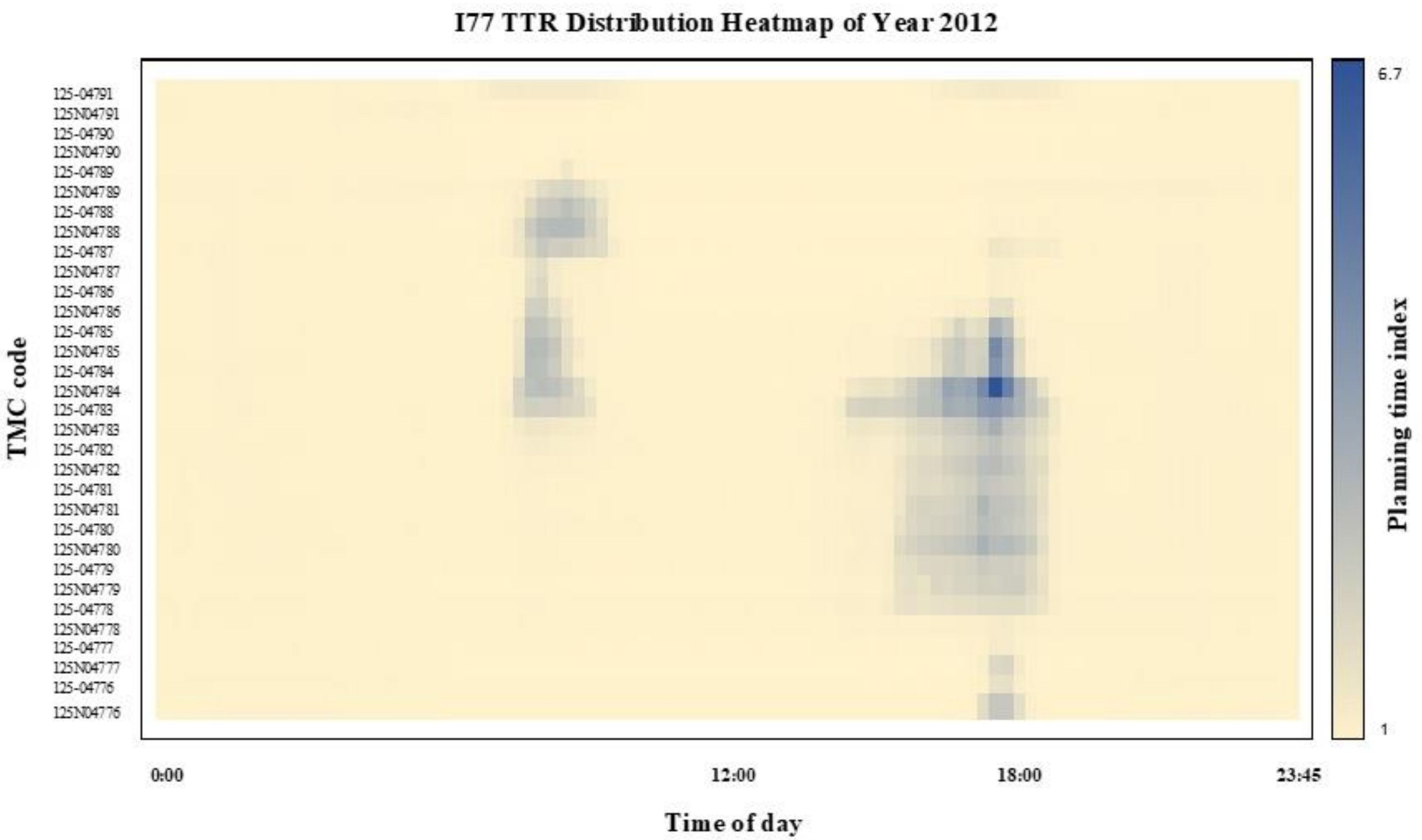


FIGURE 4.2: PTI heatmap of I77 (SB) in year 2012

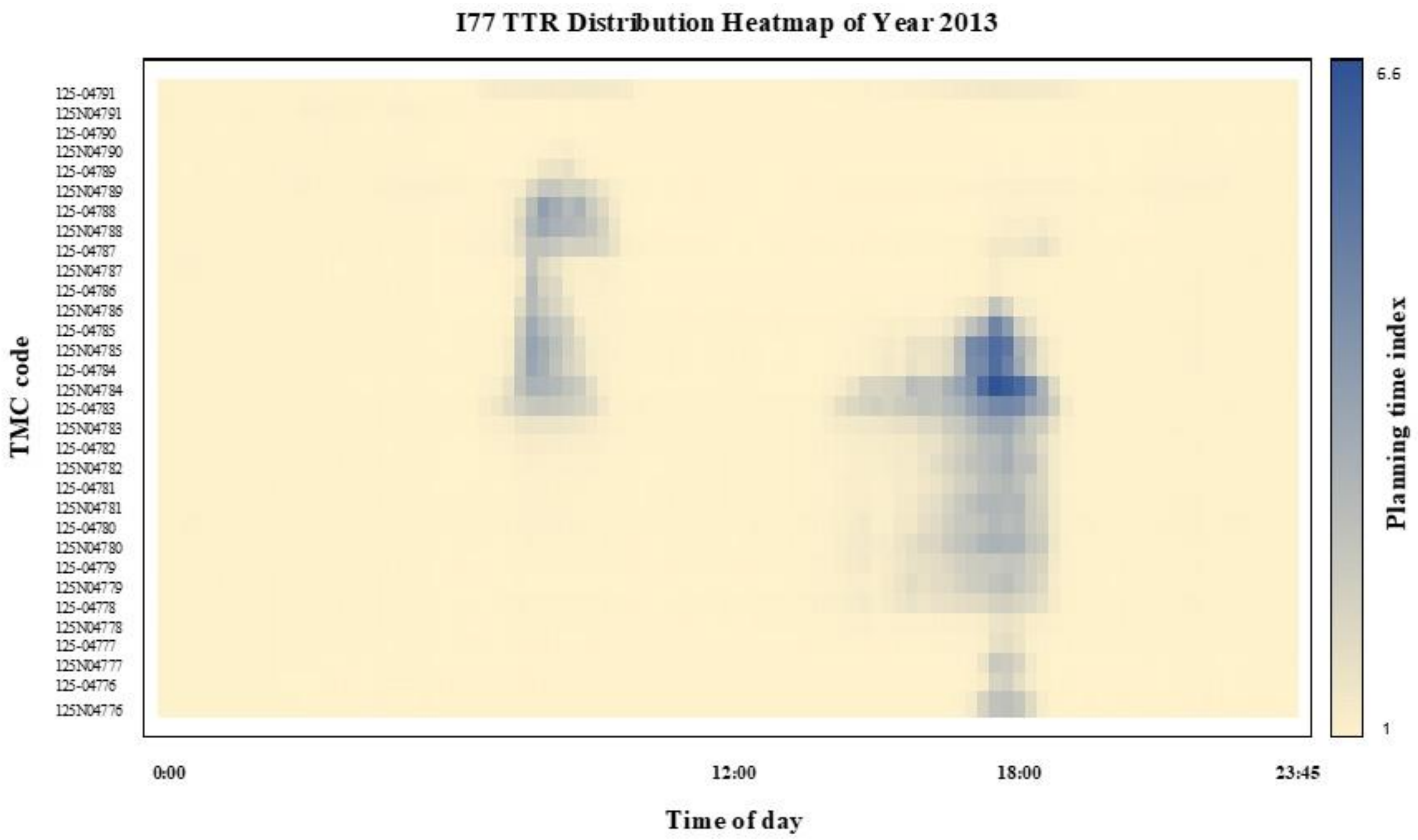


FIGURE 4.3: PTI heatmap of I77 (SB) in year 2013

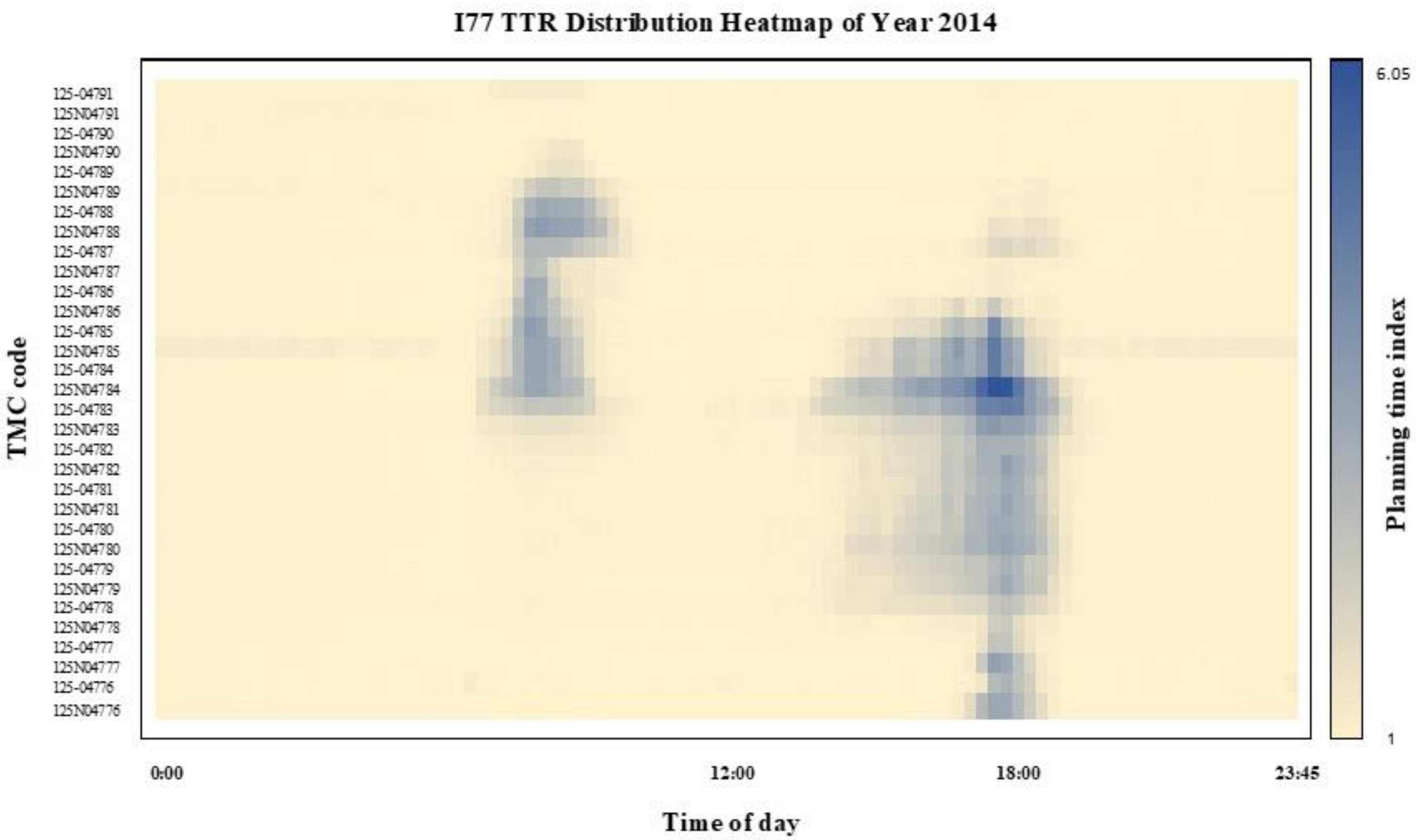


FIGURE 4.4: PTI heatmap of I77 (SB) in year 2014

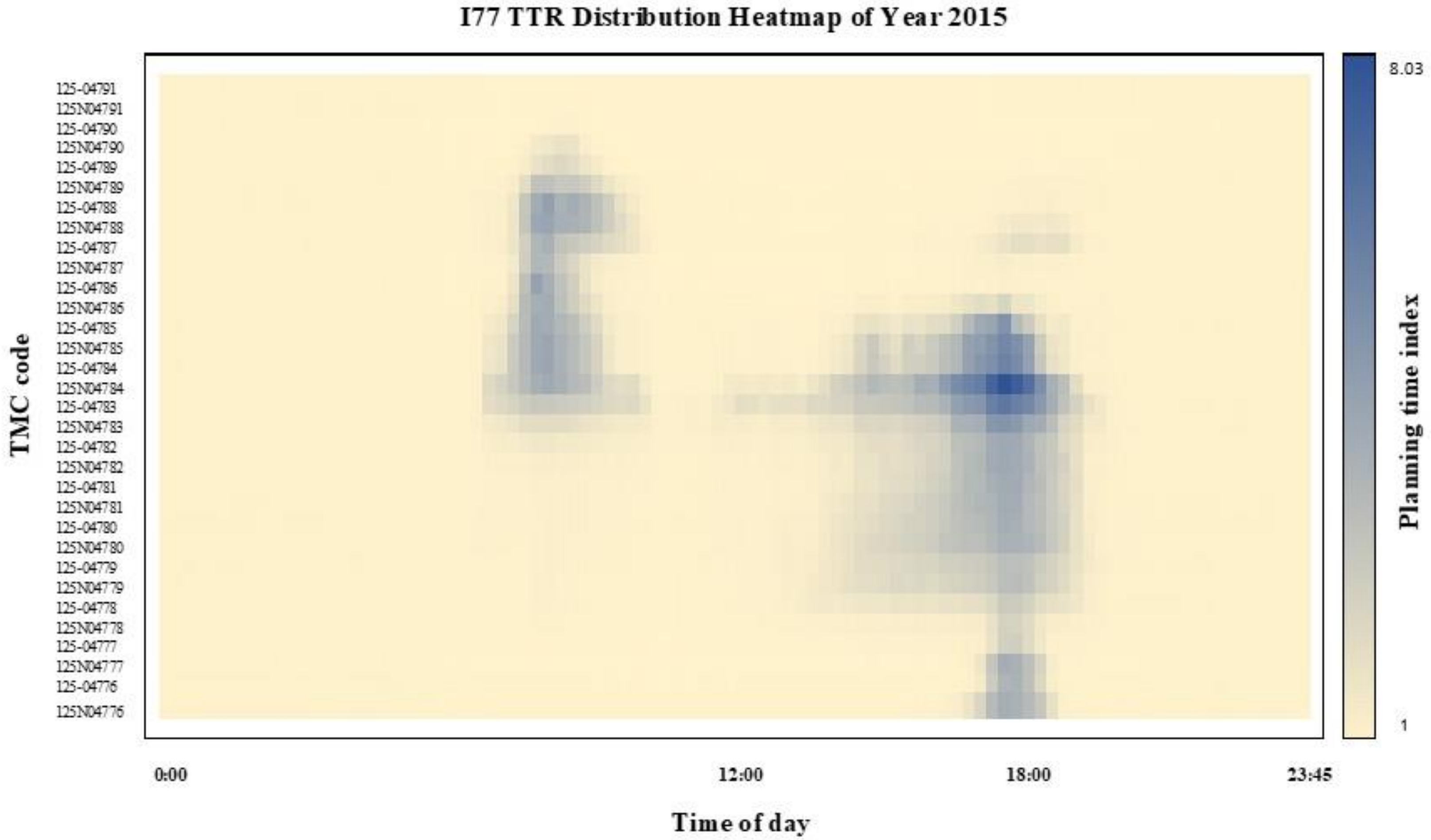


FIGURE 4.5: PTI heatmap of I77 (SB) in year 2015

The PTI heatmaps show that, during morning peak periods, traffic congestion generally occurs in the vicinity of segment 125N04783 to segment 125N04789; during evening peak periods, drivers routinely experience frequent congestion between segment 125N04776 and segment 125N04785. The study location identification criteria will be discussed in the next section.

4.2.3. Study Location Identification Based on PTI Rating

In order to select the sections which can represent different traffic conditions, the qualitative ratings for each freeway segment in the study area are conducted and further classified into different categories/levels based on the qualitative criteria of a previous study (Wolniak and Mahapatra, 2014). The ratings which are given based on the PTI values are: (1) reliable (PTI<1.5); (2) moderately to heavily unreliable (1.5<PTI<2.5) and (3) extremely unreliable (PTI>2.5).

Based on the rating criteria mentioned above, eight segments (shown in Figure 4.6) which contain the four PTI rating cases are selected as the sample study segments. The four cases are:

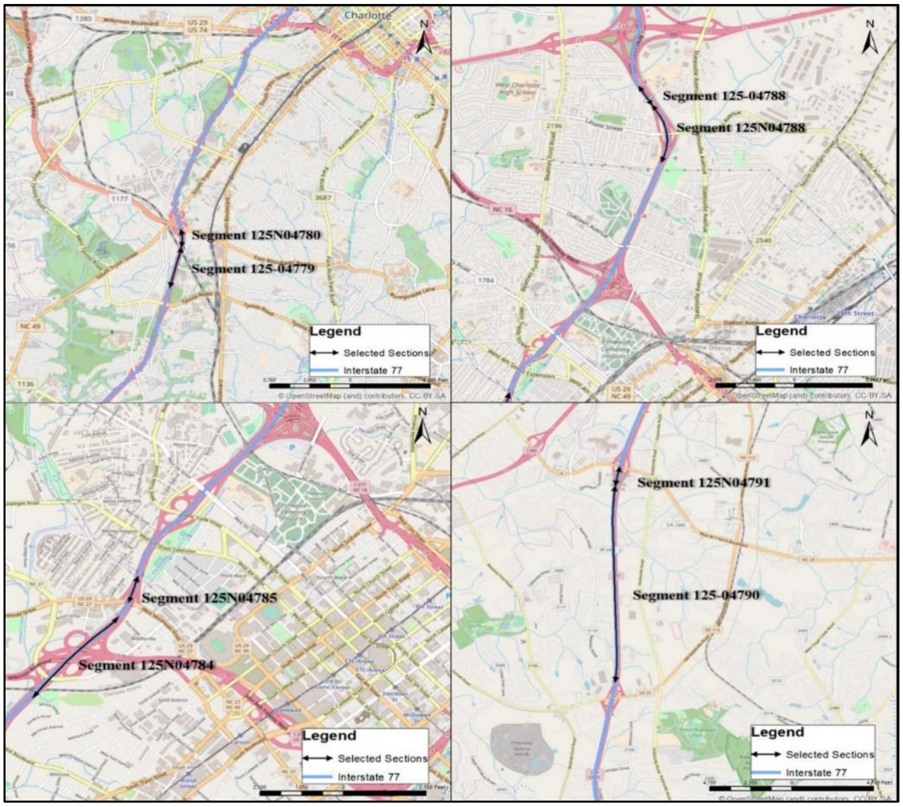


FIGURE 4.6: Locations of selected segments

Case 1 (PM peak only): The average PTI during PM peak period is reliable and during PM peak period is unreliable/extremely unreliable. The selected segments are 125-04779 and 125N04780.

Case 2 (AM peak only): The average PTI during AM peak period is unreliable/ extremely unreliable and during PM peak period is reliable. The selected segments are 125N04788 and 125-04788.

Case 3 (Double peak): The average PTI during both AM and PM peak periods are unreliable/ extremely unreliable. The selected segments are 125N04784 and 125N04785.

Case 4 (No peak): The average PTI during both AM and PM peak periods are reliable. The selected segments are 125-04790 and 125N04791.

Table 4.1 below describes the detailed information about the TMC code, segment location, segment length, year, average PTI and rating of these selected segments. The information on all the segments in the study area can be found in Appendix A.

TABLE 4.1: PTI ratings during AM and PM peak periods of selected segments

TMC Code	Segment Location	Segment Length (miles)	Year	Time Period	Average PTI	Rating
125-04779	TYVOLA RD/EXIT 5	0.67	2011	AM Peak	1.09	reliable
				PM Peak	2.00	unreliable
			2012	AM Peak	1.07	reliable
				PM Peak	1.98	unreliable
			2013	AM Peak	1.06	reliable
				PM Peak	2.08	unreliable
			2014	AM Peak	1.09	reliable
				PM Peak	2.34	unreliable
			2015	AM Peak	1.11	reliable
				PM Peak	2.70	extremely unreliable
			Average	AM Peak	1.08	reliable
				PM Peak	2.22	unreliable
125N04780	WOODLAWN RD/EXIT 6	0.26	2011	AM Peak	1.10	reliable
				PM Peak	2.45	unreliable
			2012	AM Peak	1.07	reliable
				PM Peak	2.43	unreliable
			2013	AM Peak	1.06	reliable
				PM Peak	2.49	unreliable
			2014	AM Peak	1.10	reliable
				PM Peak	2.69	extremely unreliable
			2015	AM Peak	1.12	reliable
				PM Peak	3.19	extremely unreliable
			Average	AM Peak	1.09	reliable
				PM Peak	2.65	extremely unreliable
125N04784	I-277/US-74/EXIT 9	0.94	2011	AM Peak	1.60	unreliable
				PM Peak	3.10	extremely unreliable
			2012	AM Peak	1.75	unreliable
				PM Peak	3.34	extremely unreliable

TMC Code	Segment Location	Segment Length (miles)	Year	Time Period	Average PTI	Rating
			2013	AM Peak	2.01	unreliable
				PM Peak	4.04	extremely unreliable
			2014	AM Peak	2.33	unreliable
				PM Peak	4.09	extremely unreliable
			2015	AM Peak	2.77	extremely unreliable
				PM Peak	5.45	extremely unreliable
			Average	AM Peak	2.09	unreliable
				PM Peak	4.00	extremely unreliable
125N04785	US-29/NC-27/MOREHEAD ST/EXIT 10	0.20	2011	AM Peak	1.38	reliable
				PM Peak	1.71	unreliable
			2012	AM Peak	1.60	unreliable
				PM Peak	2.08	unreliable
			2013	AM Peak	1.87	unreliable
				PM Peak	2.95	extremely unreliable
			2014	AM Peak	2.11	unreliable
				PM Peak	2.85	extremely unreliable
			2015	AM Peak	2.63	extremely unreliable
				PM Peak	3.61	extremely unreliable
			Average	AM Peak	1.92	unreliable
				PM Peak	2.64	extremely unreliable
125N04788	LASALLE ST/EXIT 12	0.53	2011	AM Peak	1.81	unreliable
				PM Peak	1.08	reliable
			2012	AM Peak	1.90	unreliable
				PM Peak	1.07	reliable
			2013	AM Peak	2.09	unreliable
				PM Peak	1.11	reliable
			2014	AM Peak	2.32	unreliable
				PM Peak	1.26	reliable
			2015	AM Peak	2.62	extremely unreliable
				PM Peak	1.25	reliable
			Average	AM Peak	2.15	unreliable
				PM Peak	1.16	reliable
125-04788		0.11	2011	AM Peak	1.72	unreliable

TMC Code	Segment Location	Segment Length (miles)	Year	Time Period	Average PTI	Rating
	LASALLE ST/EXIT 12			PM Peak	1.09	reliable
			2012	AM Peak	1.71	unreliable
				PM Peak	1.06	reliable
			2013	AM Peak	2.09	unreliable
				PM Peak	1.07	reliable
			2014	AM Peak	2.14	unreliable
				PM Peak	1.11	reliable
			2015	AM Peak	2.63	extremely unreliable
				PM Peak	1.13	reliable
			Average	AM Peak	2.06	unreliable
				PM Peak	1.09	reliable
125-04790	US-21/SUNSET RD/EXIT 16	2.25	2011	AM Peak	1.05	reliable
				PM Peak	1.05	reliable
			2012	AM Peak	1.04	reliable
				PM Peak	1.04	reliable
			2013	AM Peak	1.04	reliable
				PM Peak	1.04	reliable
			2014	AM Peak	1.06	reliable
				PM Peak	1.05	reliable
			2015	AM Peak	1.07	reliable
				PM Peak	1.05	reliable
			Average	AM Peak	1.07	reliable
				PM Peak	1.07	reliable
125N04791	HARRIS OAK BLVD/REAMES RD/EXIT 18	0.62	2011	AM Peak	1.06	reliable
				PM Peak	1.05	reliable
			2012	AM Peak	1.05	reliable
				PM Peak	1.05	reliable
			2013	AM Peak	1.07	reliable
				PM Peak	1.06	reliable
			2014	AM Peak	1.07	reliable
				PM Peak	1.07	reliable
			2015	AM Peak	1.05	reliable
				PM Peak	1.05	reliable
			Average	AM Peak	1.04	reliable
				PM Peak	1.04	reliable

4.3. Travel Time Variability Analysis at Study Locations

4.3.1. TTR Pattern under Case 1

The PTIs of segment 125-04779 and 125N04780 from 2011 to 2015 are shown in Figure 4.7 and 4.8. These two sections are located at the south part of the Charlotte downtown area. The volume of outbound traffic during PM hours is high and therefore contributes to the frequent congestion under PM peak condition. In more detail, in the year 2015, these two segments had obvious higher PTI values during peak hours than those in the years of 2011-2014. The condition like this may be attributed to different factors such as the traffic volume, weather condition and accidents. One potential reason behind this could be the traffic volume of the segments of case 1 from 2011 to 2015 (annual average daily traffic (AADT): 15300, 15200, 15400, 15900, and 15900, respectively). The correlation values between the AADT and average daily PTIs of these two segments are 0.86 and 0.83, respectively, which means highly correlated. Therefore, the traffic volume may be a primary reason of the TTR distribution characteristics. Based on the historical weather data, the frequency of adverse weather in the year 2015 is higher than that in the year from 2011 to 2014. In order to eliminate the possible influence of adverse weather, the TTR distribution under only normal conditions during each year are also tested and the average daily PTI of 2015 is reduced a little bit (from 2.1 to 2.0) but still higher than PTIs of year 2011-2014. With respect to traffic accident, no detailed historical crash information about I77 is found. However, the number of total crashes in Mecklenburg county in each year had been getting higher and higher from 2011 to 2015 (15476, 15915, 16790, 19847, and 21096, respectively) (NCDMV, 2016). This can also be another potential reason that contributes to the worsening of the traffic condition in the year 2015.

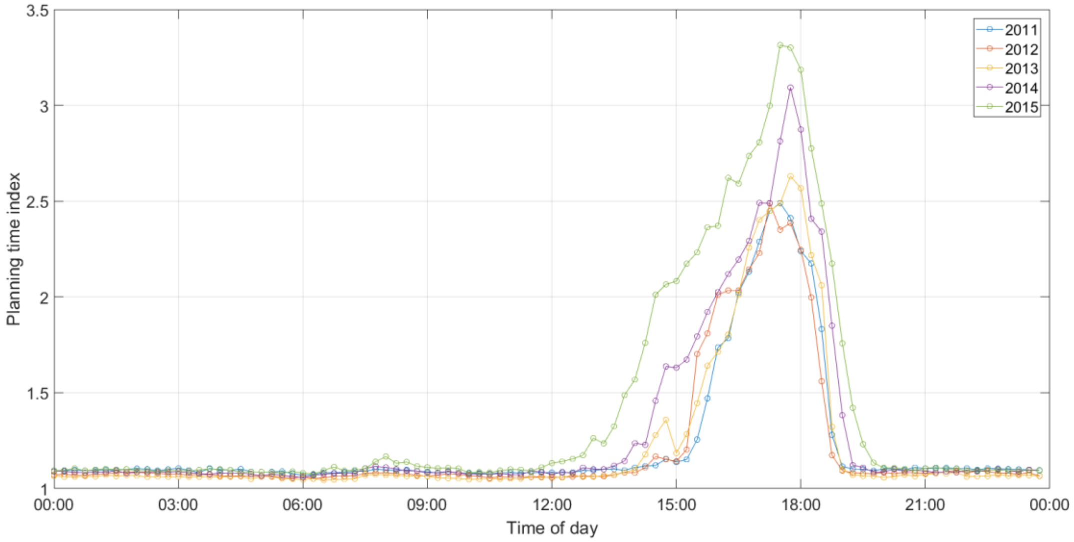


FIGURE 4.7: TTR pattern of segment 125-04779 in 5 years

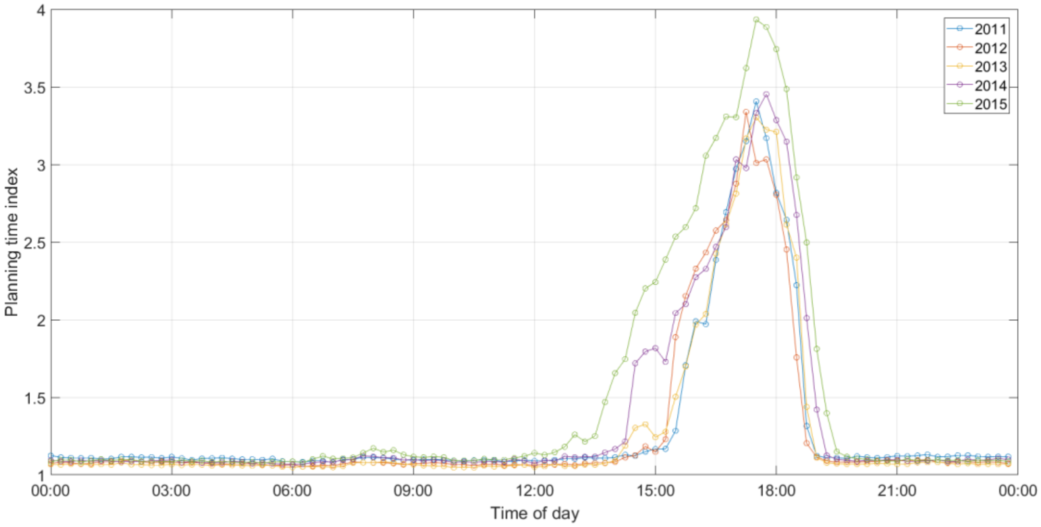


FIGURE 4.8: TTR pattern of segment 125N04780 in 5 years

4.3.2. TTR Pattern under Case 2

The PTIs of segment 125N04788 and 125-04788 from 2011 to 2015 are shown in Figure 4.9 and 4.10. These two sections are located at the north part of the Charlotte downtown area. The volume of inbound traffic during AM hours is high and therefore contributes to the frequent congestion under AM peak condition. Similar to case 1, in the year 2015, these two segments had obvious higher PTI values during peak hours than that of years of 2011-2014. The condition like this may also be explained by the potential factors (such as traffic volume (with the correlation values 0.83 and 0.89, respectively), adverse weather and accident) that contribute to the worsening of the traffic condition in the year 2015.

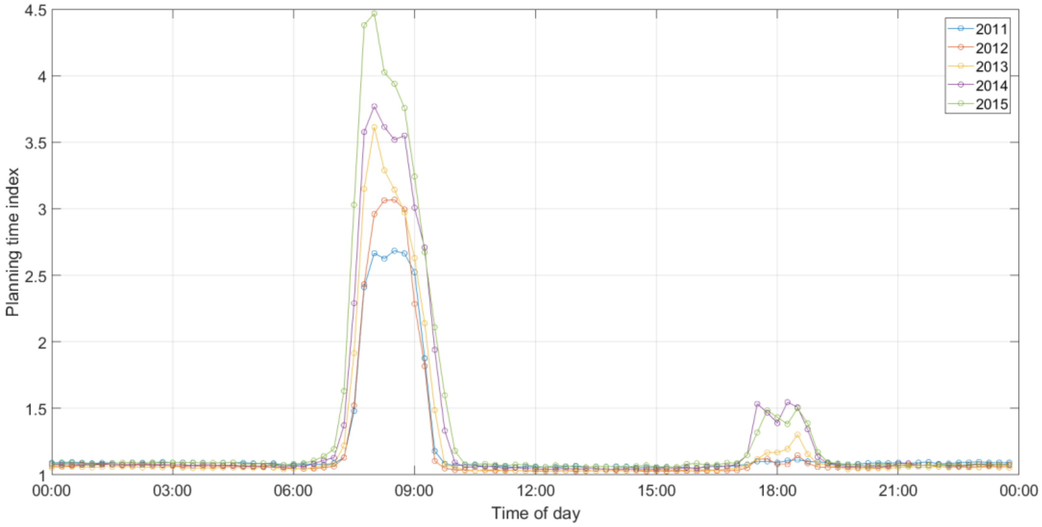


FIGURE 4.9: TTR pattern of segment 125N04788 in 5 years

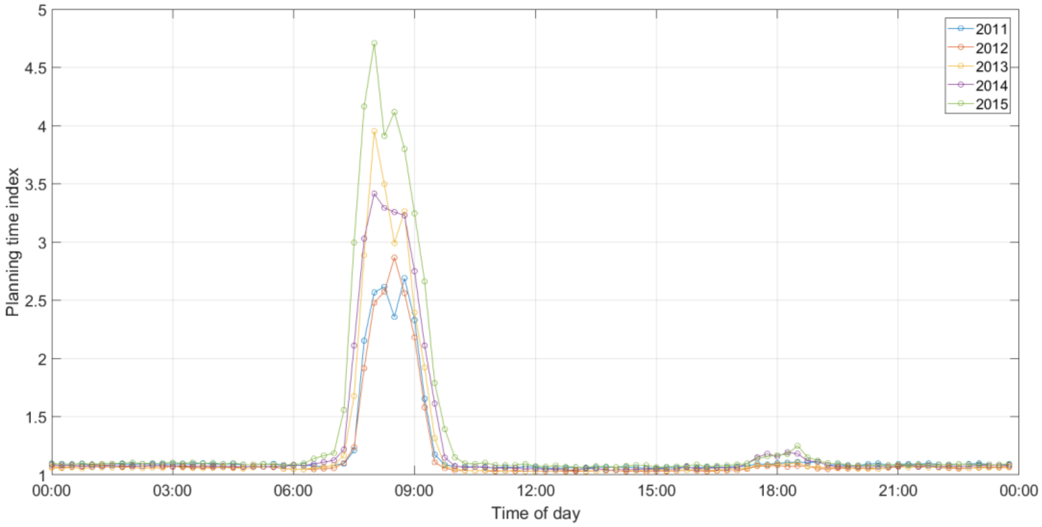


FIGURE 4.10: TTR pattern of segment 125-04788 in 5 years

4.3.3. TTR Pattern under Case 3

The PTIs of segment 125N04784 and 125N04785 from 2011 to 2015 are shown in Figure 4.11 and 4.12. These two sections are located adjacent to Charlotte downtown area. The volume of inbound traffic during AM hours and outbound traffic during PM hours are both high and therefore contributes to the frequent congestion under double peak condition. Similar to case 1 and 2, in the year 2015, these two segments had obvious higher PTI values during peak hours than those in the years of 2011-2014. However, the correlation values between traffic volume and average daily PTIs are not statistically significant (0.56 and 0.71, respectively).Therefore, the condition like this may be explained by the other potential factors (such as adverse weather and accident) that contribute to the worsening of the traffic condition in the year 2015.

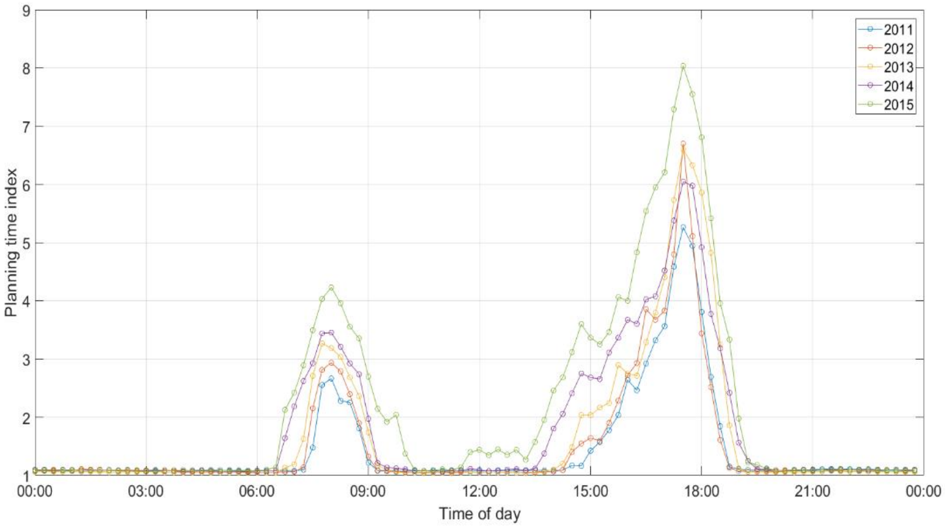


FIGURE 4.11: TTR pattern of segment 125N04784 in 5 years

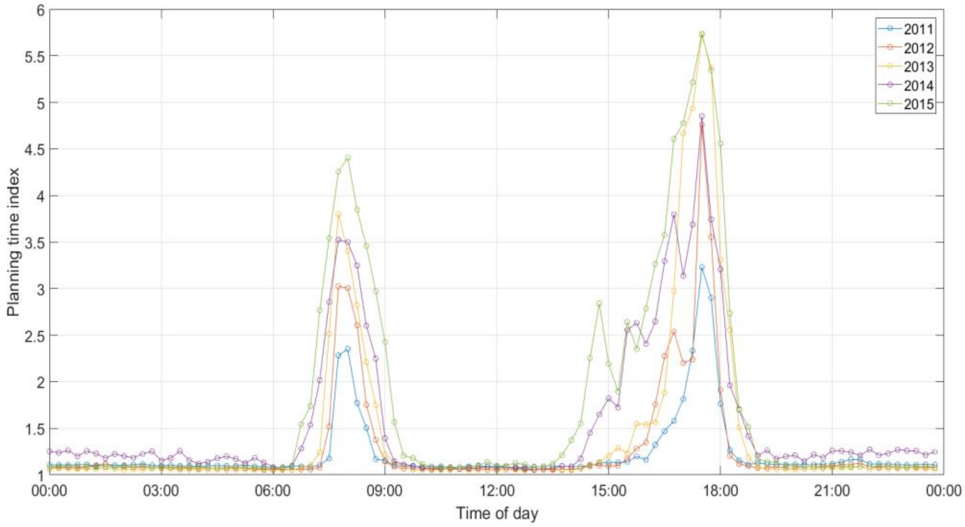


FIGURE 4.12: TTR pattern of segment 125N04785 in 5 years

4.3.4. TTR Pattern under Case 4

The PTIs of segment 125-04790 and 125N04791 from 2011 to 2015 are shown in Figure 4.13 and 4.14. These two sections are located far away from Charlotte downtown area. The traffic volumes during both AM and PM hours are low and therefore contributes to the no peak condition. The variation of PTIs throughout the day of each year do not change significantly (from 1.02 to 1.13 and 1.04 to 1.15, respectively).

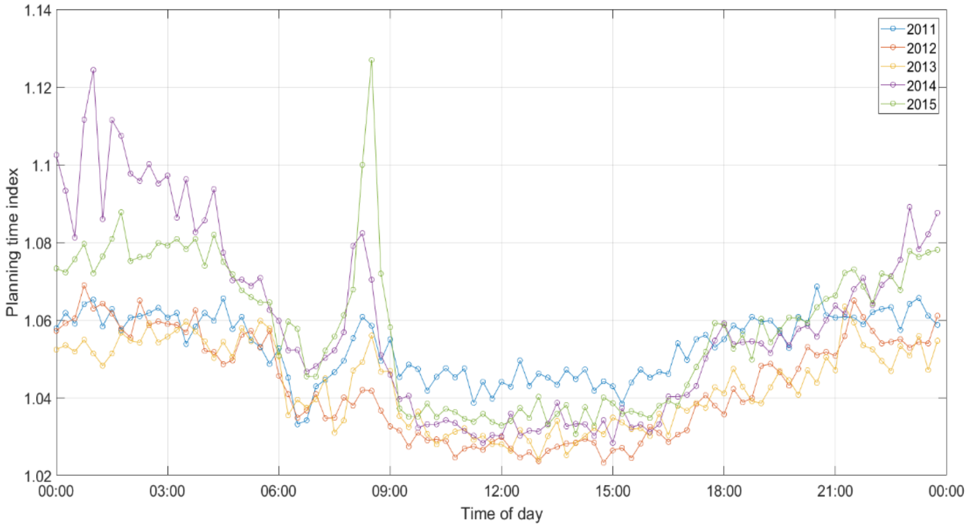


FIGURE 4.13: TTR pattern of segment 125-04790 in 5 years

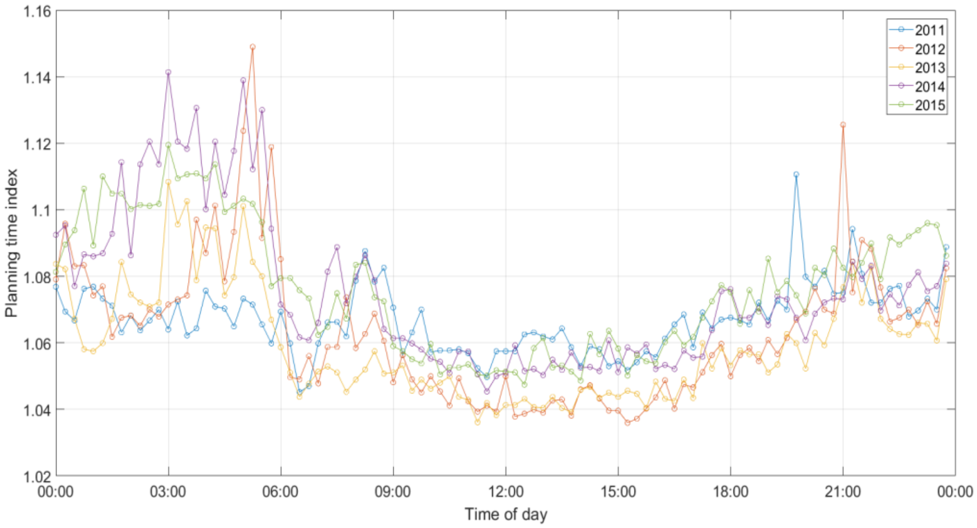


FIGURE 4.14: TTR pattern of segment 125N04791 in 5 years

4.4. Travel Time Variability Analysis of Different DOW

4.4.1. TTR Pattern of Different DOW: Case 1

The PTIs of segment 125-04779 and 125N04780 from Monday to Sunday are shown in Figure 4.15 to Figure 4.16 below, and the average PTIs are shown in Table 4.2. The PTI ranking result shows that: the TTR patterns of these two sections on weekdays are similar to the TTR pattern under all conditions. However, the TTR patterns on weekends are significantly different from weekdays. There are no PM peak characteristics of the TTR of these two segments on weekends as the PTIs throughout the day do not change significantly. The results indicate that traffic congestion on weekends becomes less frequent and also travel demand on weekends is perhaps much lower than that on weekdays, which is consistent with previous studies (Chen et al., 2017, Chen et al., 2018). The travel time on Friday is least reliable. This result is consistent with a previous study (Wang et al., 2017).

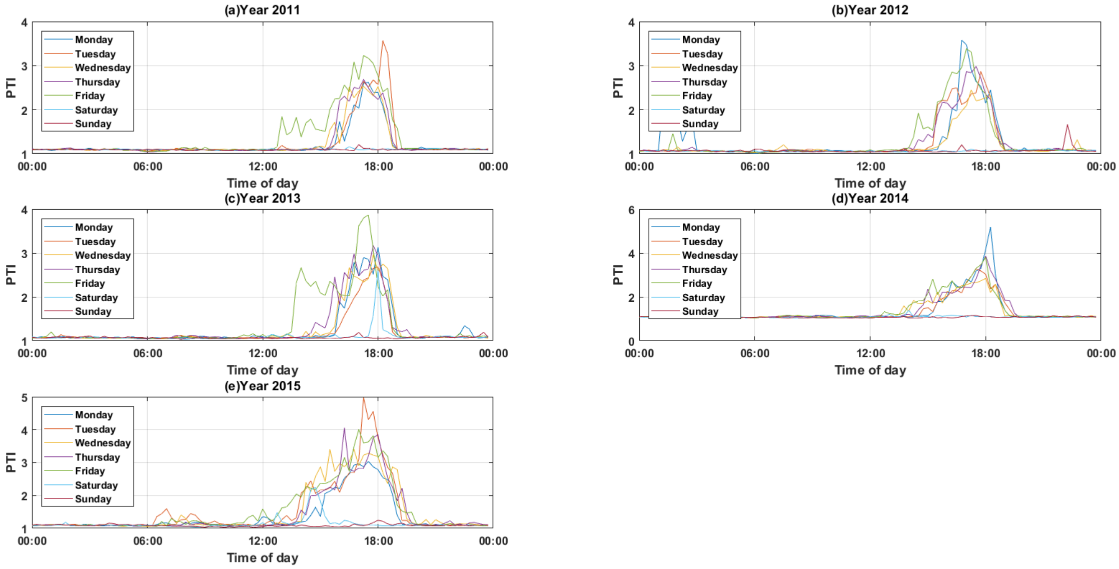


FIGURE 4.15: TTR pattern of segment 125-04779 from Monday to Sunday

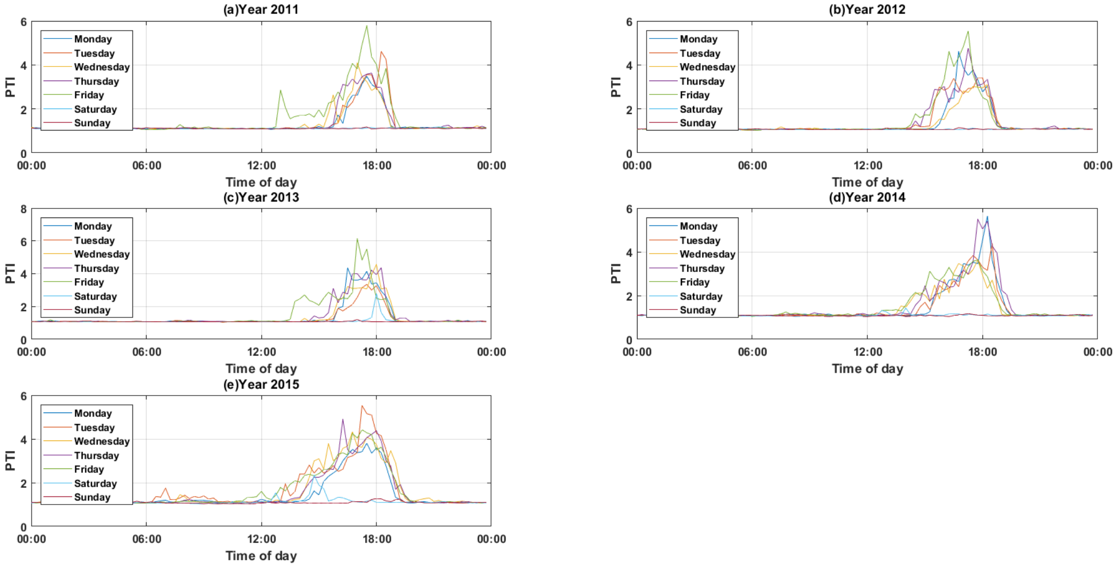


FIGURE 4.16: TTR pattern of segment 125N04780 from Monday to Sunday

TABLE 4.2: Average PTIs from Monday to Sunday (Case 1)

	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Sunday**
Segment 125-04779							
Average PTI	1.29	1.30	1.30	1.32	1.40	1.10	1.08
Rank	5	3	4	2	1	6	7
Morning Peak (7-9 am) PTI	1.10	1.14	1.11	1.10	1.11	1.06	1.06
Rank	5	1	2	4	3	6	7
Afternoon Peak (4-7 pm) PTI	2.39	2.42	2.31	2.47	2.63	1.13	1.09
Rank	4	3	5	2	1	6	7
Segment 125N04780							
Average PTI	1.37	1.39	1.38	1.44	1.51	1.11	1.09
Rank	5	3	4	2	1	6	7
Morning Peak (7-9 am) PTI	1.11	1.15	1.12	1.12	1.11	1.07	1.08
Rank	4	1	2	3	5	7	6
Afternoon Peak (4-7 pm) PTI	2.89	2.89	2.81	3.19	3.26	1.16	1.11
Rank	4	3	5	2	1	6	7

4.4.2. TTR Pattern of Different DOW: Case 2

The PTIs of segment 125N04788 and 125-04788 on different DOW are shown in Figure 4.17 to Figure 4.18 below. Similar to case 1, the TTR patterns of these two sections on weekdays are similar to the TTR patterns under all conditions and the patterns on weekends are significantly different from weekdays. There are no AM peak characteristics of the TTR of these two segments on weekends as the PTIs throughout the day do not change significantly. The results indicate that traffic congestion on weekends becomes less frequent and also travel demand of these two segments on weekends is perhaps much lower than that on weekdays. The PTI ranking result shows that the travel time on Tuesday is least reliable.

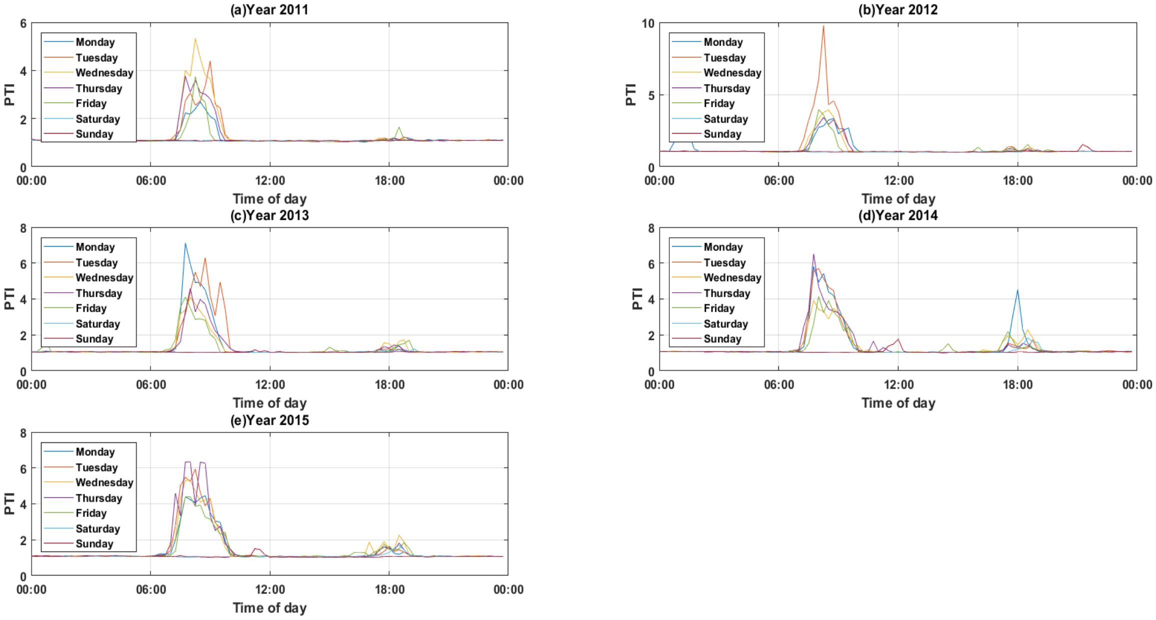


FIGURE 4.17: TTR pattern of segment 125N04788 from Monday to Sunday

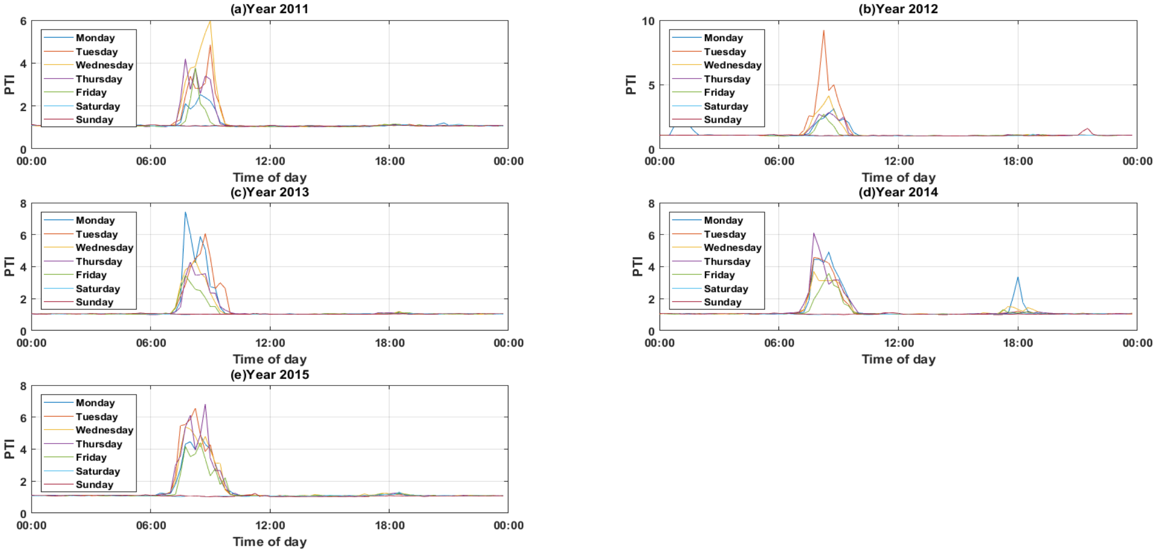


FIGURE 4.18: TTR pattern of segment 125-04788 from Monday to Sunday

TABLE 4.3: Average PTIs from Monday to Sunday (Case 2)

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Segment 125N04788							
Average PTI	1.28	1.32	1.28	1.27	1.18	1.06	1.06
Rank	3	1	2	4	5	7	6
Morning Peak (7-9 am) PTI	3.09	3.74	3.13	3.23	2.55	1.04	1.05
Rank	4	1	3	2	5	7	6
Afternoon Peak (4-7 pm) PTI	1.23	1.16	1.28	1.18	1.25	1.14	1.05
Rank	3	5	1	4	2	6	7
Segment 125-04788							
Average PTI	1.32	1.37	1.31	1.31	1.25	1.07	1.07
Rank	2	1	4	3	5	6	7
Morning Peak (7-9 am) PTI	2.91	3.53	3.11	3.01	2.14	1.04	1.05
Rank	4	1	2	3	5	7	6
Afternoon Peak (4-7 pm) PTI	1.14	1.08	1.12	1.09	1.09	1.07	1.05
Rank	1	5	2	3	4	6	7

4.4.3. TTR Pattern of Different DOW: Case 3

The PTIs of segment 125N04784 and 125-04785 on different DOW are shown in Figure 4.19 to Figure 4.20 below. Similar to case 1, the TTR patterns of these two sections on weekdays are similar to the TTR patterns under all conditions and the patterns on weekends are significantly different from weekdays. The PTIs of these two sections on weekends do not change significantly in most of the time. The unique PM peak pattern of segment 125N04784 on weekends in the year 2015 may be explained by the potential reason that higher accident rate of the year 2015. The PTI ranking result shows that the travel time on Friday is least reliable.

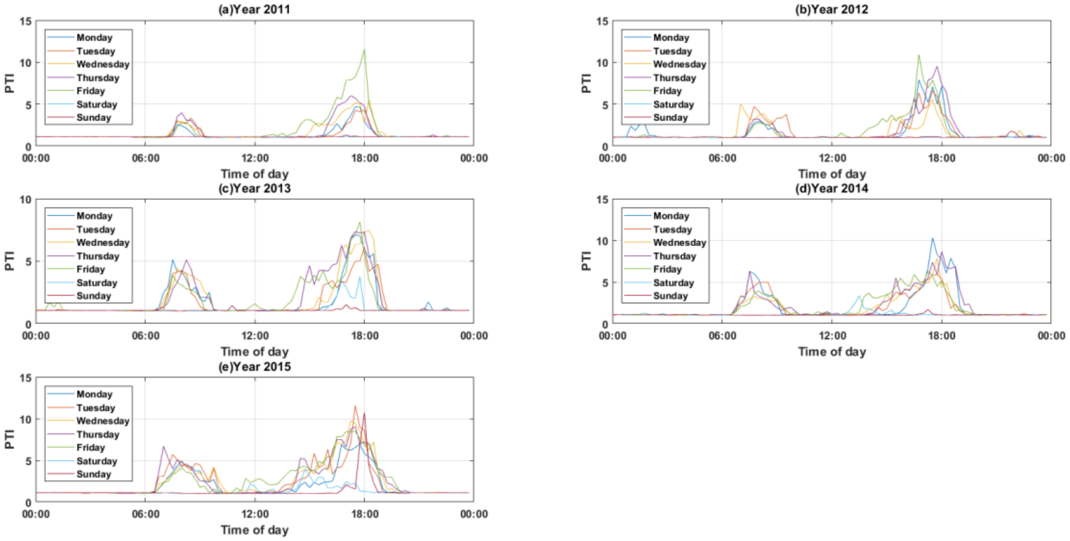


FIGURE 4.19: TTR pattern of segment 125N04784 from Monday to Sunday

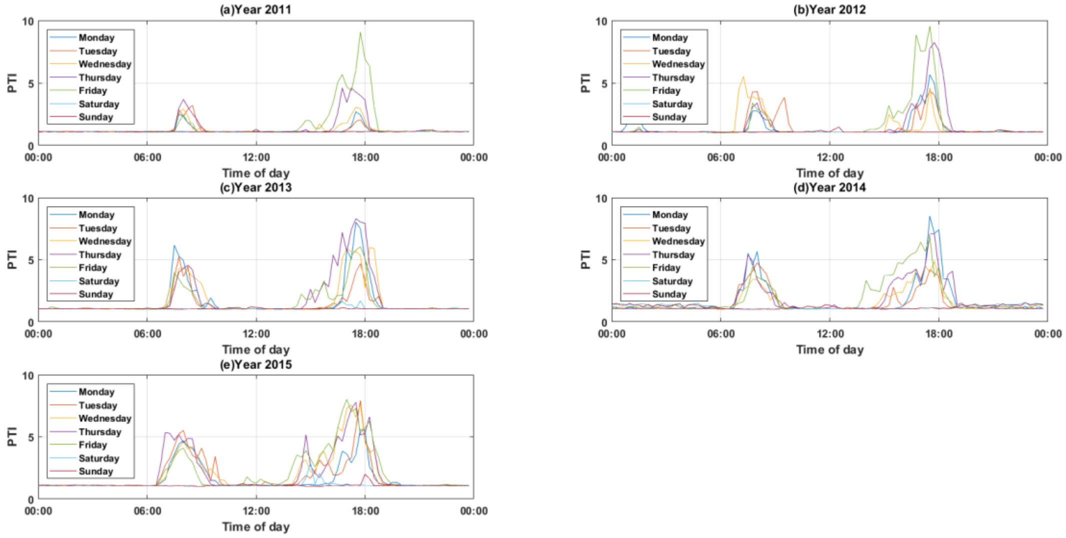


FIGURE 4.20: TTR pattern of segment 125N04785 from Monday to Sunday

TABLE 4.4: Average PTIs from Monday to Sunday (Case 3)

	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Sunday**
Segment 125N04784							
Average PTI	1.74	1.78	1.85	1.97	2.02	1.15	1.12
Rank	5	4	3	2	1	6	7
Morning Peak (7-9 am) PTI	2.98	3.31	3.09	3.08	2.49	1.05	1.05
Rank	4	1	2	3	5	6	7
Afternoon Peak (4-7 pm) PTI	4.30	3.94	4.41	5.29	5.14	1.35	1.45
Rank	4	5	3	1	2	7	6
Segment 125N04785							
Average PTI	1.49	1.46	1.57	1.73	1.77	1.11	1.11
Rank	4	5	3	2	1	6	7
Morning Peak (7-9 am) PTI	2.73	2.94	2.89	2.92	2.30	1.05	1.06
Rank	4	1	3	2	5	7	6
Afternoon Peak (4-7 pm) PTI	2.84	2.27	3.01	4.18	4.33	1.12	1.10
Rank	4	5	3	2	1	6	7

4.4.4. TTR Pattern of Different DOW: Case 4

The PTIs of segment 125-04790 and 125N04791 on different DOW are shown in Figure 4.21 to Figure 4.22 below. The PTIs of two segments during both weekdays and weekends do not change significantly. The results indicate that the travel time on these two segments do not change frequently on both weekdays and weekends.

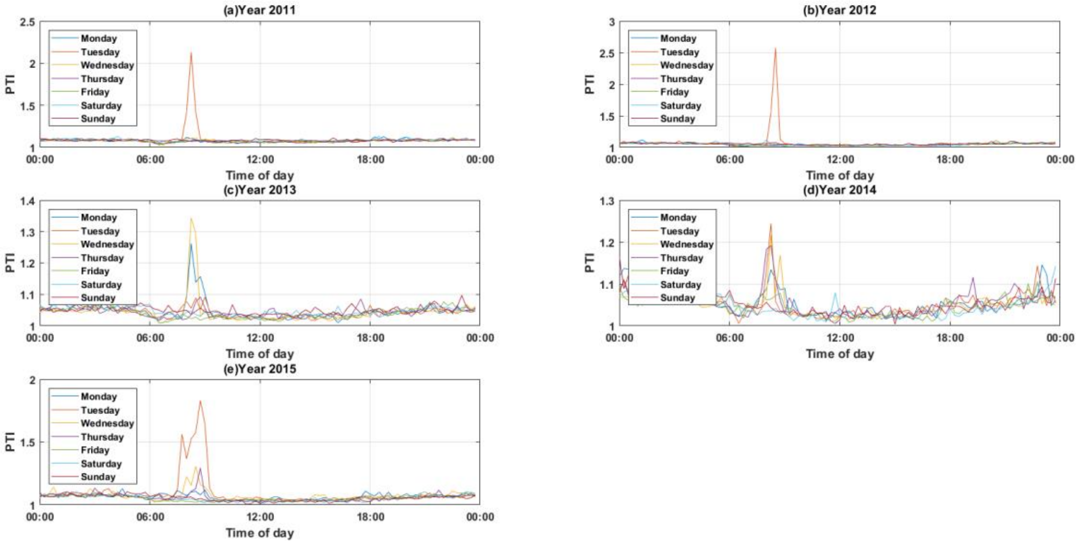


FIGURE 4.21: TTR pattern of segment 125-04790 from Monday to Sunday

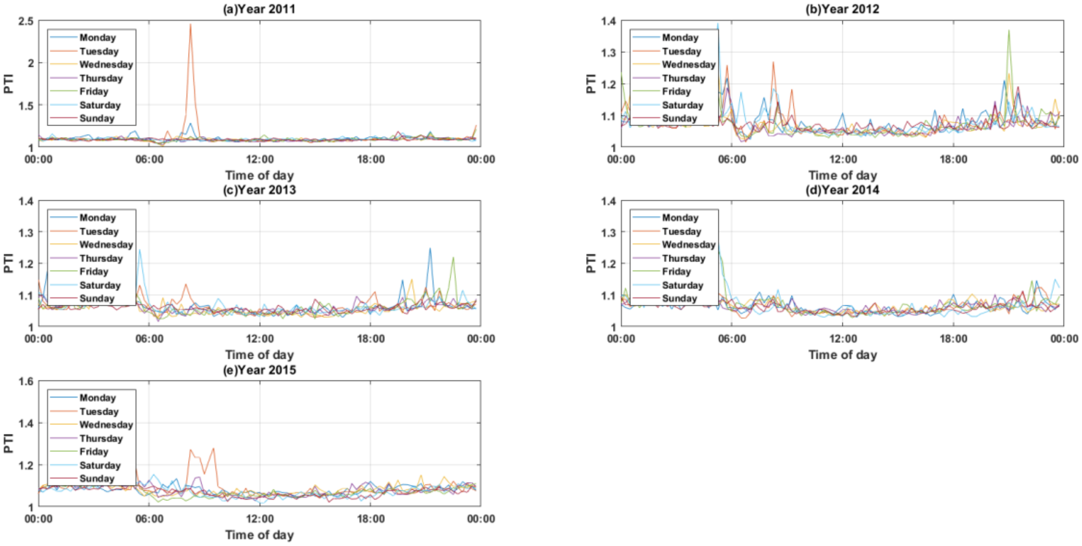


FIGURE 4.22: TTR pattern of segment 125N04791 from Monday to Sunday

TABLE 4.5: Average PTIs from Monday to Sunday (Case 4)

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Segment 125-04790							
Average PTI	1.06	1.07	1.06	1.06	1.05	1.06	1.06
Rank	3	1	4	6	7	5	2
Morning Peak (7-9 am) PTI	1.08	1.22	1.09	1.07	1.04	1.05	1.06
Rank	3	1	2	4	7	6	5
Afternoon Peak (4-7 pm) PTI	1.05	1.05	1.05	1.05	1.05	1.04	1.05
Rank	3	1	4	6	5	7	2
Segment 125N04791							
Average PTI	1.08	1.09	1.07	1.08	1.08	1.08	1.07
Rank	2	1	6	5	4	3	7
Morning Peak (7-9 am) PTI	1.09	1.14	1.07	1.07	1.06	1.08	1.07
Rank	2	1	5	4	7	3	6
Afternoon Peak (4-7 pm) PTI	1.07	1.07	1.07	1.07	1.07	1.06	1.06
Rank	2	1	5	3	4	6	7

4.5. Travel Time Variability Analysis under Different Weather Conditions

4.5.1. TTR Pattern of Different Weather Conditions: Case 1

The PTIs of segment 125-04779 and 125N04780 under different weather conditions are shown in Figure 4.23 and Figure 4.24 below. The TTR patterns of these two sections under normal and rain conditions are similar and the pattern is unique under the snow/ice/fog condition. In more detail, the PTIs under rain condition have obvious higher values than normal condition throughout the day. This probably suggests that rain can cause several travel problems such as visibility issues while driving a vehicle. Heavy rainfall may lead to hydroplaning, slippery surfaces for tires and road flooding. Therefore, the values of PTIs under rain condition also increase and the traffic congestion becomes more frequent. This result is consistent with other studies (Tsapakis et al. 2013, Li et al. 2016). The PTIs under snow/ice/fog condition is also higher than those under normal condition throughout the day because of the influence of road surfaces and visibility problems (Weng et al., 2013). The potential reason for the unique TTR pattern under the snow/fog/ice condition could be: snow/fog/ice can contribute to unexpected condition on the roadway anytime throughout the day. This result is also consistent with a previous study (Yazici et al., 2011). In specific, there is an extremely high PTI value at noon. Since the geometric design characteristics of all the segments are similar, the potential reason behind this unique pattern could be the non-recurrent condition such as the incidents happened during snow condition at the case segments. This hypothesis should be checked in the future if more detailed data is available.

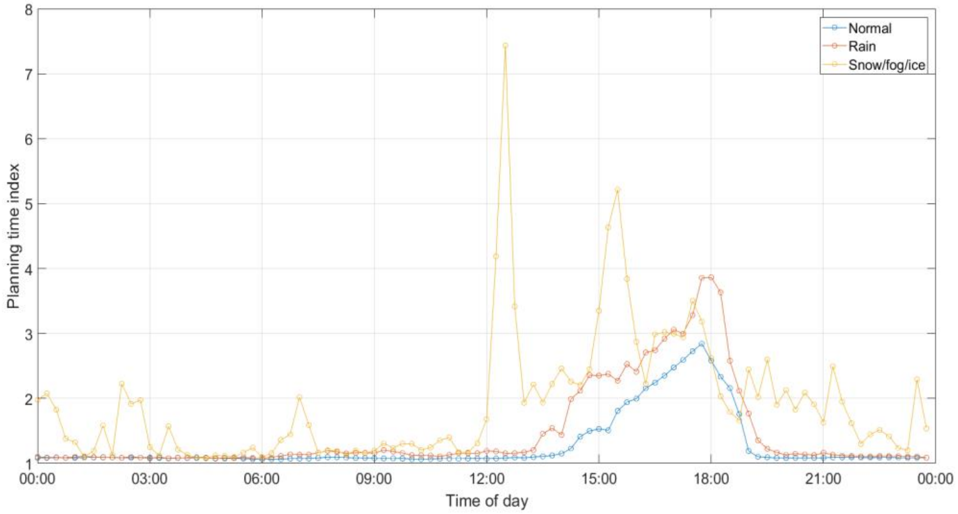


FIGURE 4.23: TTR pattern of segment 125-04779 under different weather conditions

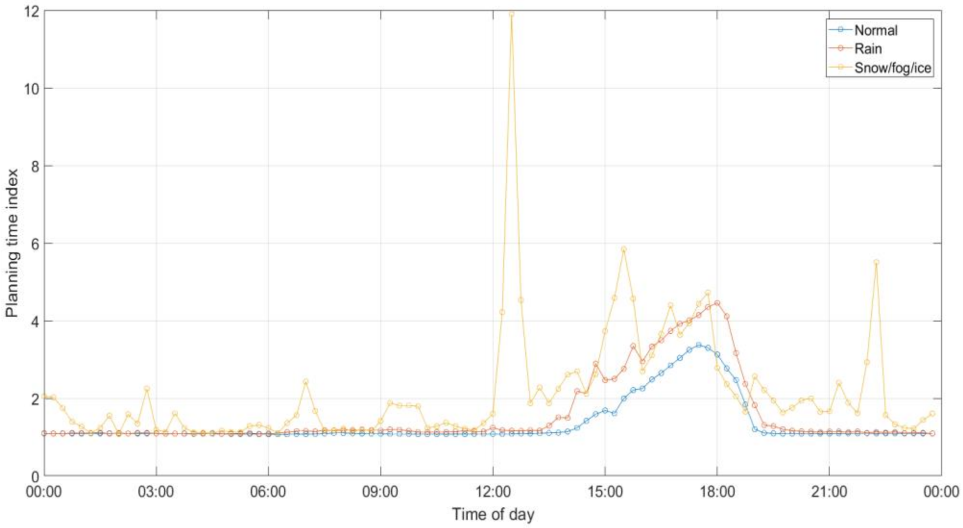


FIGURE 4.24: TTR pattern of segment 125N04780 under different weather conditions

4.5.2. TTR Pattern of Different Weather Conditions: Case 2

The PTIs of segment 125N04788 and 125-04788 under different weather conditions are shown in Figure 4.25 and Figure 4.26 below. Similar to case 1, the PTIs under rain condition have obvious higher values than those under normal condition throughout the day. And the PTIs under the snow/ice/fog condition are also higher than the PTIs under normal condition throughout the day and demonstrates unique variability pattern.

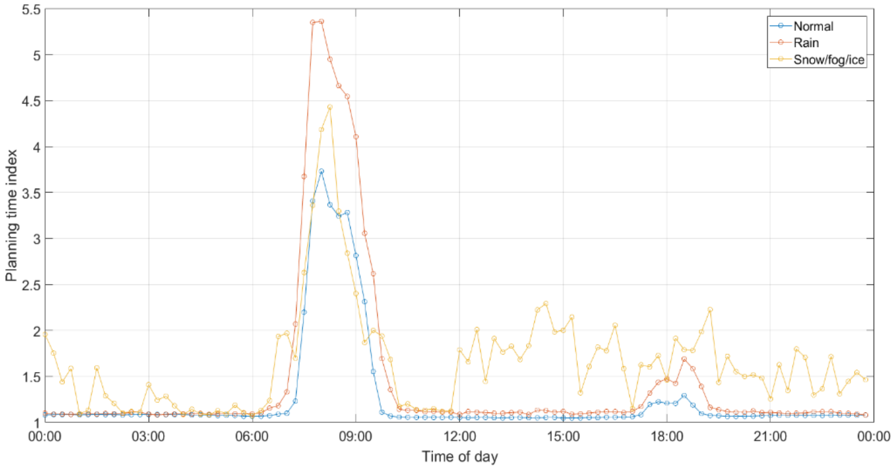


FIGURE 4.25: TTR pattern of segment 125N04788 under different weather conditions

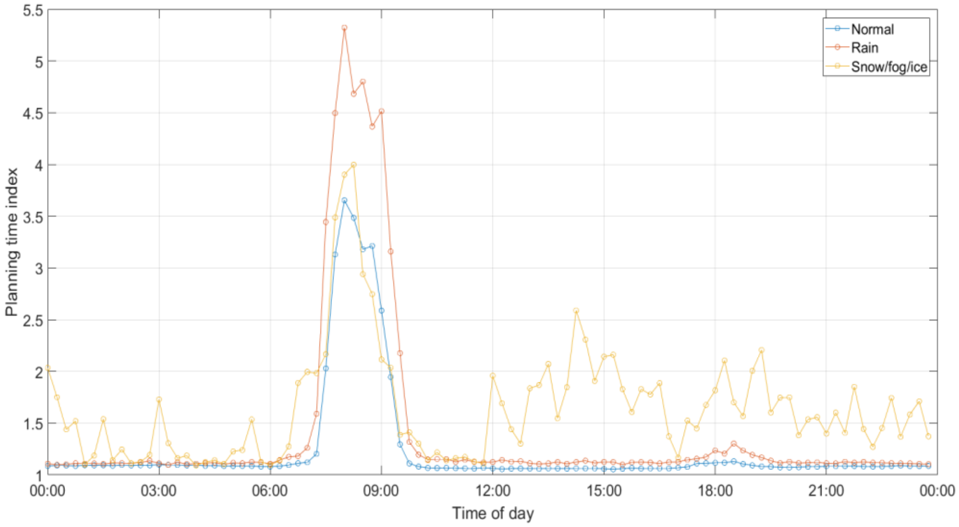


FIGURE 4.26: TTR pattern of segment 125-04788 under different weather conditions

4.5.3. TTR Pattern of Different Weather Conditions: Case 3

The PTIs of segment 125N04784 and 125N04785 under different weather conditions are shown in Figure 4.27 and Figure 4.28 below. Similar to case 1 and 2, the PTIs under rain condition have obvious higher values than those under normal condition throughout the day. And the PTIs under the snow/ice/fog condition is also higher than the PTIs under normal condition throughout the day and demonstrates unique variability pattern.

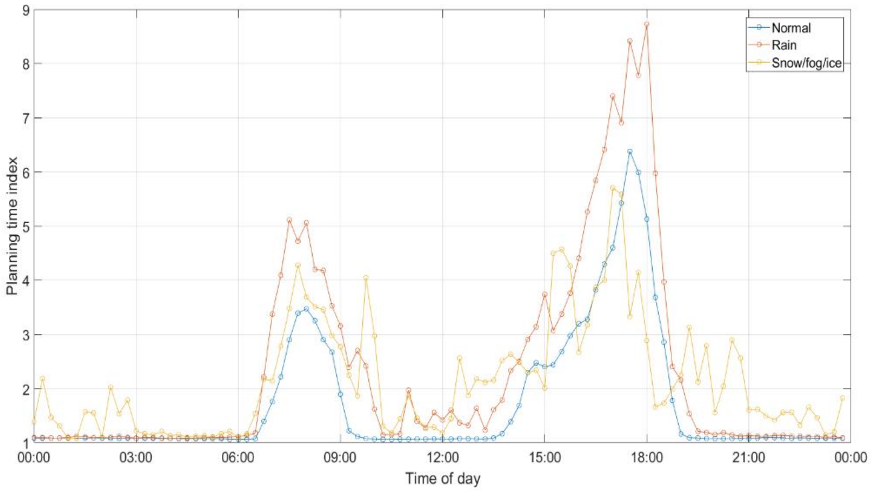


FIGURE 4.27: TTR pattern of segment 125N04784 under different weather conditions

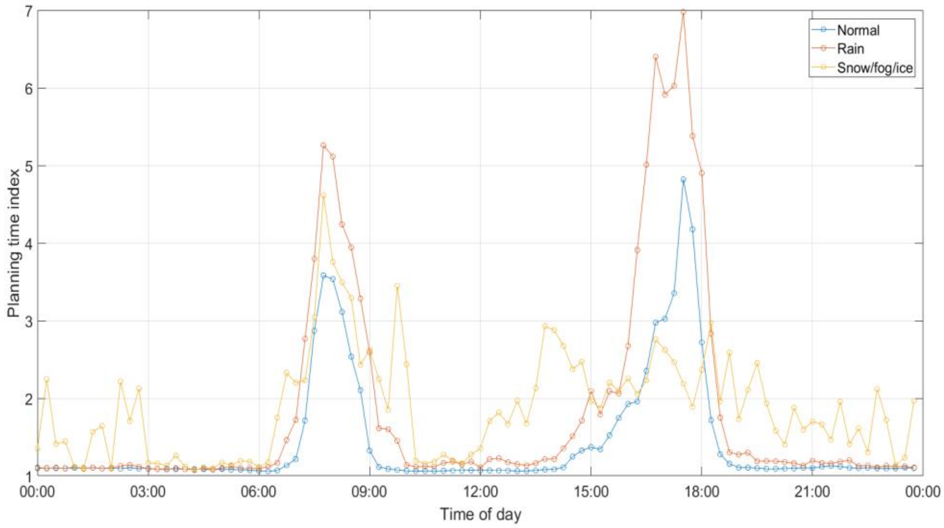


FIGURE 4.28: TTR pattern of segment 125N04785 under different weather conditions

4.5.4. TTR Pattern of Different Weather Conditions: Case 4

The PTIs of segment 125-04790 and 125N04791 under different weather conditions are shown in Figure 4.29 and Figure 4.30 below. In more detail, the PTIs under rain condition have higher values than normal condition but not increase significantly. However, the PTIs under the snow/ice/fog condition is much higher than the PTIs under normal condition throughout the day. This result shows the adverse weather like snow, fog and ice can affect the traffic condition of the segment significantly, and the traffic congestion becomes more frequent no matter when.

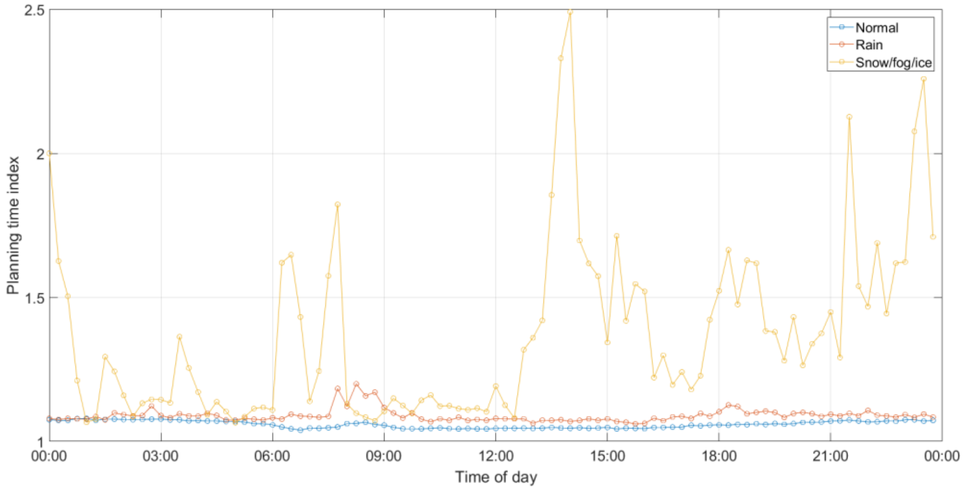


FIGURE 4.29: TTR pattern of segment 125-04790 under different weather conditions

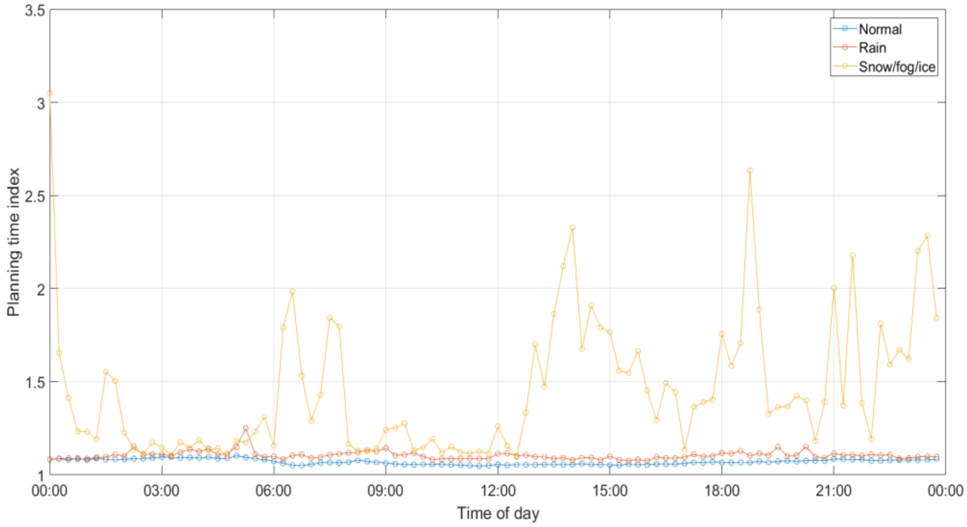


FIGURE 4.30: TTR pattern of segment 125N04791 under different weather conditions

4.6. Summary

This chapter describes the analysis results of travel time variability patterns. The analysis results can give a clear picture of the travel time variability characteristics under general condition, on different DOW and under different weather conditions.

CHAPTER 5: TRAVEL TIME PREDICTION METHODOLOGY

5.1. Introduction

This chapter presents the introduction to the travel time prediction methodology. The following sections are organized as follows. Section 5.2 shows the basic information about the ensemble learning methodology, which includes the ideas of bagging algorithm and boosting algorithm. Section 5.3 discusses the principles of the XGBoost algorithm. Finally, section 5.4 concludes this chapter with a summary.

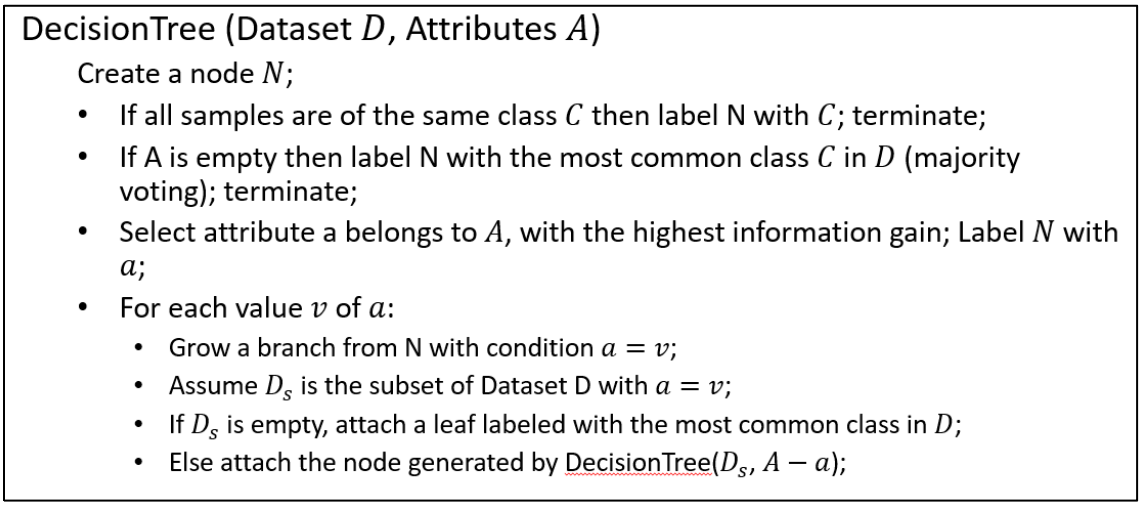
5.2. Basic Information on the Ensemble Learning Methodology

The ensemble learning-based algorithms consist of multiple base models (e.g., decision tree model), and each base model provides an alternative solution to the problem. The prediction results of these base models are combined by some rules (such as weighted or unweighted voting and averaging). The final output will be achieved through the combined model (Zhang and Haghani, 2015).

The base model of the ensemble learning algorithm is extremely important to the final results. Since the model is expected to have enough degrees of freedom to solve the underlying complexity of the data and avoid high variance and be more robust at the same time, the two most fundamental characteristics of the base model should be a low bias and a low variance. In other words, the base model should be a ‘weak learner’ and needs to be converted to a ‘strong learner’. In machine learning area, a ‘weak learner’ is a model which performs relatively poorly but better than random guessing.

Decision tree is a basic data-driven supervised learning method and has been widely used in the data mining area (Quinlan, 1986; Han and Kamber, 2011). A single decision tree is constructed by splitting the features' space into regions. The target variable can be predicted by using the values of a set of features.

In detail, the pseudo-code for decision tree is shown below in Figure 5.1, which can make it easier to understand the idea of decision tree algorithm.



```
DecisionTree (Dataset $D$, Attributes $A$)
    Create a node $N$;
    • If all samples are of the same class $C$ then label N with $C$; terminate;
    • If A is empty then label N with the most common class $C$ in $D$ (majority
      voting); terminate;
    • Select attribute a belongs to $A$, with the highest information gain; Label $N$ with
      $a$;
    • For each value $v$ of $a$:
        • Grow a branch from N with condition $a = v$;
        • Assume $D_s$ is the subset of Dataset D with $a = v$;
        • If $D_s$ is empty, attach a leaf labeled with the most common class in $D$;
        • Else attach the node generated by DecisionTree($D_s$, $A - a$);
```

FIGURE 5.1: Pseudo-code for decision tree (Source: Quinlan 1986)

Tree model is a commonly used base models in ensemble learning. Tree model can be very sensitive, and the computation process of tree model is fast and easy, which can reduce model complexity and improve the efficiency.

Overfitting means that a function fits the data too well. Typically, this is because the actual equation is too complicated to consider each data point and outlier. The tree-based ensemble method can reduce the variance by building a large number of trees and then combining the results of them.

The purpose of an ensemble learning algorithm is to achieve an improved result by combining predictions of a group of individual base models. It has been shown that the combined model often generates more stable and accurate predictions in many applications (Leblanc and Tibshirani, 1996; Banfield et al., 2006).

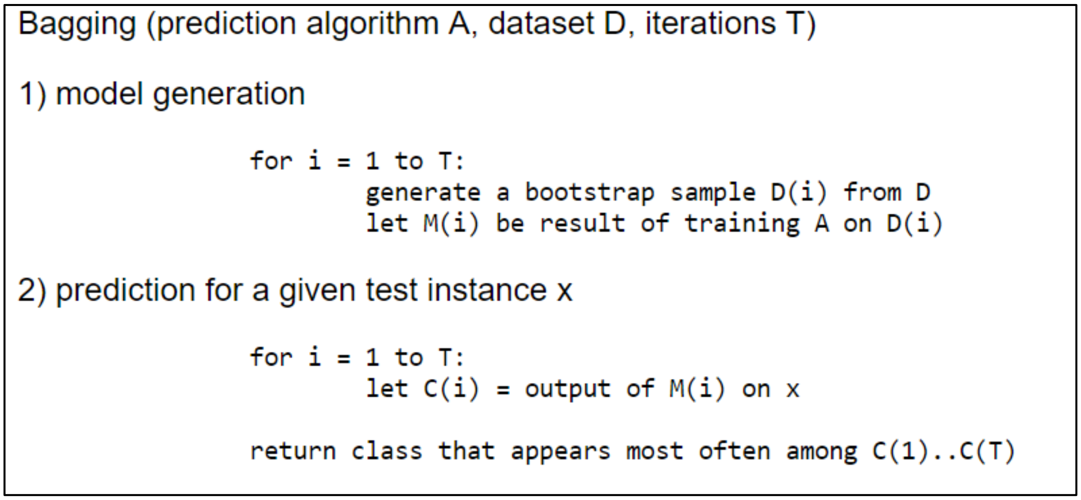
Bagging and boosting are both ensemble techniques, where a set of base models are combined to create a model that obtains better performance than a single model. However, they utilize different re-sampling methods and therefore can have different performances and generate different outputs.

5.2.1. Bagging algorithm

Bagging is a method for generating multiple versions of predictor and using these to get an aggregated predictor (Breiman, 1996). The bagging algorithm could help reduce the overfitting problem from a single model.

Typically, there are 3 steps to use the tree-based bagging algorithm: The first step is to create several (e.g., 100) random sub-samples of the dataset with replacement. The second step is to train a model using each sample. Finally, given a new dataset, calculate the average prediction from each model (Breiman, 1996).

In detail, the pseudo-code for bagging introduced by Breiman (1996) is shown in Figure 5.2, which can make it easier to understand the idea of bagging algorithm. Given a training set D, in each iteration (ranges from 1 to T), randomly sample with replacement N samples from the training dataset. Then train a selected base model A (e.g., decision tree model) on samples. For each test example, start with all trained base models, and then predict by combining results of all T trained models. For the regression problem, the combining rule will be averaging them; for the classification problem, the combining rule will be a majority vote.



```
Bagging (prediction algorithm A, dataset D, iterations T)

1) model generation

        for i = 1 to T:
            generate a bootstrap sample D(i) from D
            let M(i) be result of training A on D(i)

2) prediction for a given test instance x

        for i = 1 to T:
            let C(i) = output of M(i) on x

        return class that appears most often among C(1)..C(T)
```

FIGURE 5.2: Pseudo-code for bagging (Source: Breiman, 1996)

Random Forest is a typical bagging-based model that was introduced by Breiman (2001), and it has been widely used in the machine learning area. Random Forest is a combination of many decision trees. There are two types of randomness built into the trees. First, each tree is built on a random sample from the training dataset. Second, a subset of features are randomly allocated to each tree node to generate the best split.

The main limitation of the Random Forest is that a larger number of trees may make the model run slower. If the data include categorical variables with a different number of levels, "*Random Forests are biased in favor of those variables with more levels*" (Strickland, 2007).

5.2.2. Boosting algorithm

The idea of boosting algorithm was first proposed by Kearns (1988). Boosting algorithm also refers to several algorithms that convert weak learners to strong learners. Several base models are combined together to form stronger model that can make generalizations (Rajsingh et al., 2018).

Different from the bagging method which has each base model run independently and then aggregates their outputs at the end without any preference, the boosting method improves the prediction through developing multiple models in sequence by putting emphasis on these training cases that are difficult to estimate.

In detail, the initial model in boosting is predicted using a loss function. Each time a decision tree is generated, the model is updated based on the previous model and loss function resulting in a final model. The samples with the highest error appear most in subsequent models, which means that the incorrectly estimated or misclassified samples have more chances to be selected.

There are many boosting algorithms such as AdaBoost, Gradient boosting, and XGBoost. Gradient boosting is a typical boosting approach, and it has been widely used in the machine learning area. The word ‘gradient’ means that it uses a gradient descent algorithm to minimize the loss when adding new models (Friedman, 2001). The gradient boosting approach supports both classification and regression predictive modeling problems.

Based on previous studies, the gradient boosting model generally gives better results than Random Forest, since Random Forest has fewer parameters needing tuning and also is less sensitive to these parameters (Ogutu et al., 2011; Freeman et al., 2015). However, the gradient boosting model is harder to fit than Random Forests at the same time. The stopping criteria should also be chosen carefully to avoid overfitting on the training data.

5.3. XGBoost Algorithm

XGBoost is the short name for ‘Extreme gradient boosting’ that was proposed by Chen and Guestrin (2016). In recent years, it has a recognized impact in solving machine learning challenges in different application domains.

The speed of XGBoost is much faster than that of other common machine learning methods since it can process large amounts of data in a parallel way efficiently. The XGBoost model can also handle missing values in the dataset. Above all, “*XGBoost used a more regularized model formalization to control over-fitting, which gives it better performance*” (Chen and Guestrin, 2016). Therefore, the XGBoost model is selected and used to conduct travel time prediction in this study. The detailed information about the XGBoost model is described as follows:

The objective function (Obj(Θ)) of the XGBoost model is provided below (Chen and Guestrin, 2016):

$$Obj(\Theta) = L(\Theta) + \Omega(\Theta)$$

where,

$L(\Theta)$ = The training loss, which measures how well the model fit on training data

$\Omega(\Theta)$ = The regularization term, which measures the complexity of the model.

The loss on training data can be expressed as:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i)$$

In detail, the square loss for the regression problem can be expressed as:

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$$

The logistic loss for the classification problem can be expressed as:

$$l(y_i, \hat{y}_i) = y_i \ln\left(1 + e^{-\hat{y}_i}\right) + (1 - y_i) ln\left(1 + e^{\hat{y}_i}\right)$$

In this study,

$\hat{y}_i$ = the predicted travel time.

y_i = the actual travel time.

When a new tree is added to the model, the objective function can be transformed to:

$$Obj(\mathrm{t}) = \sum_{i=1}^{n} l\big(y_i, \hat{y}_i{}^{(t)}\big) + \sum_{i=1}^{t} \Omega(f_i)$$

$$= \sum_{i=1}^{n} l\big(y_i, \hat{y}_i{}^{(t-1)} + f_t(x_i)\big) + \Omega(f_t) + constant$$

In order to get the simplest goal, the constant term should be removed from the function. The process of XGBoost uses second order Taylor expansion to extend the loss function and removes the constant term (Chen and Guestrin, 2016).

$$Obj(\mathrm{t}) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i{}^{(t-1)} + g_i f_t(x_i) + \frac{1}{2} h_i f_t{}^2(x_i)\right) + \Omega(f_t)$$

where,

$g_i = \partial_{\hat{y}_i{}^{(t-1)}} l(y_i, \hat{y}_i{}^{(t-1)})$, which means the first order partial derivative of the function

$h_i = \partial^2_{\hat{y}_i{}^{(t-1)}} l(y_i, \hat{y}_i{}^{(t-1)})$, which means the second order partial derivative of the function

After the removal of all the constants, the specific objective at step t becomes:

$$Obj(\mathrm{t}) = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t{}^2(x_i)] + \Omega(f_t)$$

In the XGBoost model, the complexity is defined as (Chen and Guestrin, 2016):

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

where,

T = the number of leaf nodes

γ = the penalty coefficient of the number of leaves

λ = the penalty coefficient of regularization

w_j= the score of leaf j

After re-formulating the tree model, the objective function with the t-th tree can be written as:

$$Obj(\mathrm{t}) = \sum\nolimits_{i=1}^{n} [g_i w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

$$Obj(\mathrm{t}) = \sum\nolimits_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T$$

where $I_j = \{i | q(x_i) = j\}$ is an instance set assigned to the j-th leaf. The objective function could be further compressed as:

$$Obj(\mathrm{t}) = \sum\nolimits_{j=1}^{T} \left[G_j w_j + \frac{1}{2}\left(H_j + \lambda\right) w_j^2 \right] + \gamma T$$

where $G_j = \sum_{i \in I_j} g_i$, $H_j = \sum_{i \in I_j} h_i$

The best w_j one can get for the objective function is:

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

Therefore, the final objective function can be written as:

$$Obj(\mathrm{t}) = -\frac{1}{2} \sum\nolimits_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T$$

The smaller the score is, the better the structure is.

XGBoost can also add branches for each leaf node. The loss reduction after the split can be expressed as:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma$$

where $\frac{G_L^2}{H_L+\lambda}$ is the score of the left node after the cut. $\frac{G_R^2}{H_R+\lambda}$ is the score of the right node after the cut. $\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}$ is the score of combination without the cut. Finally, the best structure of the model can be obtained which can minimize the objective function by enumerating different kinds of tree structures.

5.4. Summary

This chapter presents the methodology which will be used in travel time prediction. The idea of ensemble learning is introduced first. The detailed information on the decision tree algorithm, bagging algorithm, and boosting algorithm is then presented. The basic information about the Random Forest model and the gradient boosting model is also introduced. The advantages and disadvantages of each model are discussed. The basic information about the XGBoost model is also presented in this chapter. The advantages of the XGBoost model are listed. The detailed process of the XGBoost model is described including the objective function, regularization terms, and model score.

CHAPTER 6: TRAVEL TIME PREDICTION MODEL VALIDATION

6.1. Introduction

This chapter presents the validation of the proposed machine learning model based on the data described in Chapter 3. Section 6.2 shows the feature selection and pre-processing steps, the features will include but not limit to: time of day, day of the week, month, weather conditions, and segment characteristics. Section 6.3 introduces the parameters in the model and discusses the parameters' tuning process. Finally, section 6.4 concludes the chapter with a summary.

6.2. Feature Selection and Pre-processing

Determining which feature to use in the model is the most important factor of a successful machine learning algorithm (Domingos, 2012). The definition of feature engineering is "*an act of extracting features from raw data and transforming them into the formats that are suitable for the machine learning model*" (Zheng and Casari, 2018). Therefore, the quality of the features will have great influence on whether the travel time prediction model is good or not.

The real-world travel time data provided by the RITIS website (which was mentioned in chapter 3) are used for this study. The quality of the data is precise enough with less than a 0.5% missing rate (4348 out of 906048). Therefore, this study simply replaces the missing values with the mean of its closest surrounding values.

Based on previous studies (Min and Wynter, 2011; Wang et al., 2018), the features that influence the accuracy of travel time prediction may not only include the basic features (such as time of day, day of the week, month, and weather), but also include the spatial and temporal characteristics of the segments. Therefore, the travel time information from several steps before and the travel time information of adjacent segments are also selected and will be used in the model.

Table 6.1 summarizes the basic information on the features used in this study and Table 6.2 is an example of the dataset. In Table 6.2, the first 19 columns are the input variables that are used to predict travel time at time step t and the last column is the travel time. In some cases, the target variable will be transformed when it is not normally distributed, however, *“since regression tree is the basic learner of XGBoost, there is no need to normalize samples, which means that features from different units would not affect the prediction result”* (Dong et al. 2018).

For the Categorical Variable, the most commonly used method is One-hot encoding in the Python software. One-hot encoding is a process by which categorical variables are converted into a form that could be provided to machine learning algorithms to do a better job in prediction. For example, the category weekdays with 7 variables will be transferred as dummy variables. It should be noticed that if the range of category variable is too large (over hundreds of variables), this method is not suitable anymore.

TABLE 6.1: Definitions and attributes on selected features

Variable	Definition	Attribute
ID	Segment ID	Categorical
L	Length of the segment	Categorical
TOD	The TOD is represented by every 15-minute timestep indexed from 1 to 96	Categorical
DOW	The DOW is indexed from 1 to 7 to represent from Monday through Sunday	Categorical
Month	The Month is indexed from 1 to 12 to represent January to December	Categorical
Weather	Weather is indexed from 1 to 3 to represent normal, rain and snow/ice/fog, respectively	Categorical

T_{t-1}	Travel time at time step t-1 (15 minutes before)	Float
T_{t-2}	Travel time at time step t-2 (30 minutes before)	Float
T_{t-3}	Travel time at time step t-3 (45 minutes before)	Float
ΔT_{t-1}	Travel time change value at time step t-1 (15 minutes before)	Float
ΔT_{t-2}	Travel time change value at time step t-2 (30 minutes before)	Float
ΔT_{t-3}	Travel time change value at time step t-3 (45 minutes before)	Float
T_{t-1}^{i-1}	Travel time of first upstream segment at time step t-1 (15 minutes before)	Float
T_{t-1}^{i-2}	Travel time of second upstream segment at time step t-1 (15 minutes before)	Float
ΔT_{t-1}^{i-1}	Travel time change value of first upstream segment at time step t-1 (15 minutes before)	Float
ΔT_{t-1}^{i-2}	Travel time change value of second upstream segment at time step t-1 (15 minutes before)	Float
T_{t-1}^{i+1}	Travel time of first downstream segment at time step t-1 (15 minutes before)	Float
T_{t-1}^{i+2}	Travel time of second downstream segment at time step t-1 (15 minutes before)	Float
ΔT_{t-1}^{i+1}	Travel time change value of first downstream segment at time step t-1 (15 minutes before)	Float
ΔT_{t-1}^{i+2}	Travel time change value of second downstream segment at time step t-1 (15 minutes before)	Float
T_t	Travel time at time step t	Float

TABLE 6.2: Example of the raw inputs of the model

ID	Weather	TOD	DOW	Month	L	T_{t-1}	T_{t-2}	T_{t-3}	...	ΔT_{t-3}	T_{t-1}^{i-1}	T_{t-1}^{i-2}	ΔT_{t-1}^{i-1}	ΔT_{t-1}^{i-2}	T_{t-1}^{i+1}	T_{t-1}^{i+2}	ΔT_{t-1}^{i+1}	ΔT_{t-1}^{i+2}	T_t
125-04790	normal	21	Friday	1	2.25	121.91	121.54	127.10	...	-8.19	33.87	29.47	-0.31	-0.39	30.63	91.23	-0.21	1.24	116.28
125-04790	normal	22	Friday	1	2.25	116.28	121.91	121.54	...	5.56	34.70	30.88	-0.83	-1.41	29.41	90.87	1.22	0.36	117.98
125-04790	normal	23	Friday	1	2.25	117.98	116.28	121.91	...	-0.37	34.35	30.12	0.35	0.76	28.75	86.16	0.66	4.71	113.31
125-04790	normal	24	Friday	1	2.25	113.31	117.98	116.28	...	5.63	31.54	27.34	2.81	2.78	27.71	83.33	1.04	2.83	111.28
125-04790	normal	25	Friday	1	2.25	111.28	113.31	117.98	...	-1.70	31.33	26.36	0.21	0.98	27.50	82.45	0.21	0.88	108.89
125-04790	normal	26	Friday	1	2.25	108.89	111.28	113.31	...	4.67	30.58	26.38	0.75	-0.02	27.20	83.74	0.30	-1.29	118.92

6.3. Parameter Tuning Process

In the XGBoost model, there are many parameters that should be considered. There are three types of parameters: general parameters, booster parameters and task parameters.

General parameters are related to which booster is being used to do boosting, commonly in the tree or linear models. In detail, the general parameters include:

- **Booster:** Select the type of model to run at each iteration. It has 2 options: tree-based models and linear models. The default value of booster is '*gbtree*'.
- **Silent:** Silent controls whether to print message. If the value is set to 1, no running messages will be printed. The default value of silent is 0. It is generally good to keep it as 0 since the messages might help in understanding the model.
- **Nthread:** This parameter is used for controlling the parallel processing and the number of cores in the system that would be used. The default value is the maximum number of threads available on the computer. The algorithm will detect it automatically.

Booster parameters depend on which booster one has chosen. For the tree booster in this study, the parameters include:

- **Learning rate:** Learning rate is the rate at which the model learns patterns in data. After every round, it shrinks the feature weights to reach the best optimum. Lower learning rate leads to slower computation. The default value is 0.3.
- **Gamma:** Gamma controls regularization (or prevents overfitting). The optimal value of gamma depends on the data set and other parameter values. The larger the gamma is, the more conservative the algorithm will be. The value of Gamma usually is 0. The default value is 0.

- **Max_depth:** Maximum depth controls the depth of the tree. The larger the depth, the more complex the model, and the higher chance of overfitting. There is no standard value for max_depth. Larger dataset requires deeper tree to learn the rules from data. The value of Max_depth usually ranges from 3 to 10. The default value is 6.
- **Min_child_weight:** Minimum child weight refers to the minimum number of instances required in a child node. It blocks the potential feature interactions to prevent overfitting. The default value is 1.
- **Subsample:** Percentage of samples used per tree. This parameter will also help to prevent overfitting. The value of subsample usually ranges from 0.5 to 1. The default value is 1.
- **Colsample_bytree:** Percentage of features used per tree. A high value can lead to overfitting. The value of colsample_bytree usually ranges from 0.5 to 1. The default value is 1.
- **Lambda:** This parameter can help to handle the regularization part of the XGBoost model. Usually, the value of Lambda is 1 and the default value is 1.
- **Alpha:** This parameter can also help to handle the regularization part of the XGBoost model. The value of Alpha usually is 0 and the default value is 0.
- **N_estimators:** This parameter refers to the number of trees one wants to build in the model. The number is up to the complexity of the model.

Task parameters depend on the learning scenario. For example, regression tasks may use different parameters with ranking tasks. The task parameters include:

- **Objective:** This parameter defines the task of learning (the loss function to be minimized). The mostly used values are ‘reg:linear’, ‘binary:logistic’, ‘multi:softmax’ and ‘multi:softprob’. The default value is ‘reg:linear’.

In order to optimize the modelling result, it is necessary to explore the effect of different combinations of parameters on the model performance. Based on previous studies (Zhang and Haghani, 2015; Dong et al. 2018), the parameters that could be optimized include, but are not limited to: N_estimators (number of trees), learning rate, and Max_depth (maximum depth of the tree). Therefore, these parameters are considered to be optimized in this study.

There are several optimization methods considered in previous studies and the grid search method is the most widely used one. Therefore, the grid search method is selected as the optimization method with the consideration of time-efficiency. In this study, 80% of the traffic data is used as training data and 20% of the data is used as the testing data. The XGBoost model is fitted with various number of trees (N_estimators ranges from 1 to 500), maximum depth (Max_depth ranges from 5 to 10) and learning rates (Learning_rate ranges from 0.1 to 0.5). The number of stopping rounds is set as 50, which means stopping iteration after 50 rounds when there is no performance improvement.

Figure 6.1 to Figure 6.6 below show the effects of different selected variables on the prediction results. Table 6.3 below presents the detailed prediction results including the prediction results at each step, computation time, and optimized results. The mean absolute error (MAE) is used to evaluate the performance of the model.

The equation of the MAE is provided below:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \widehat{y_i}|$$

where,

m = The total number of the data.

y_i = The actual travel time value in the test dataset of record i.

$\hat{y}_i$= The predicted travel time value in the test dataset of record i.

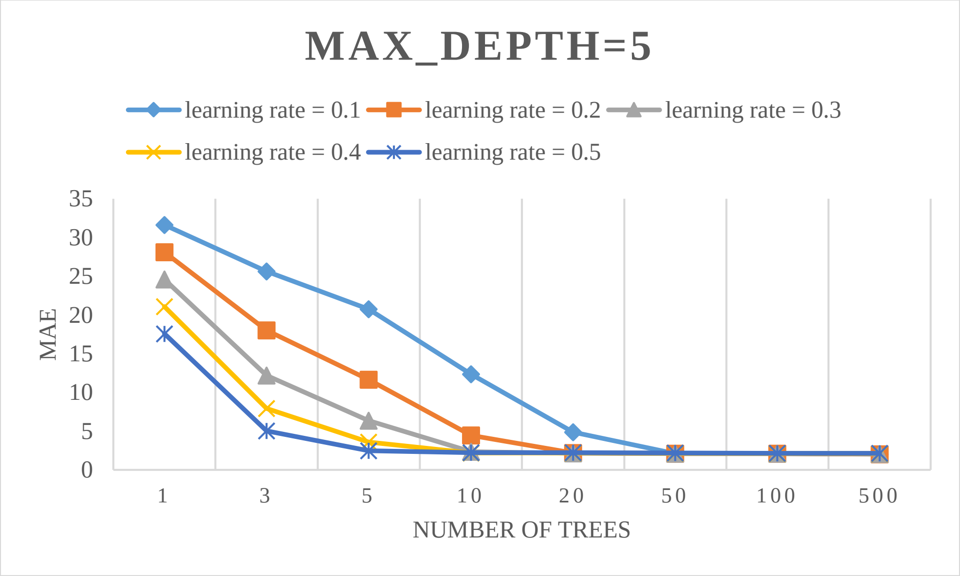


FIGURE 6.1: XGBoost travel time prediction model outputs with the Max_depth =5

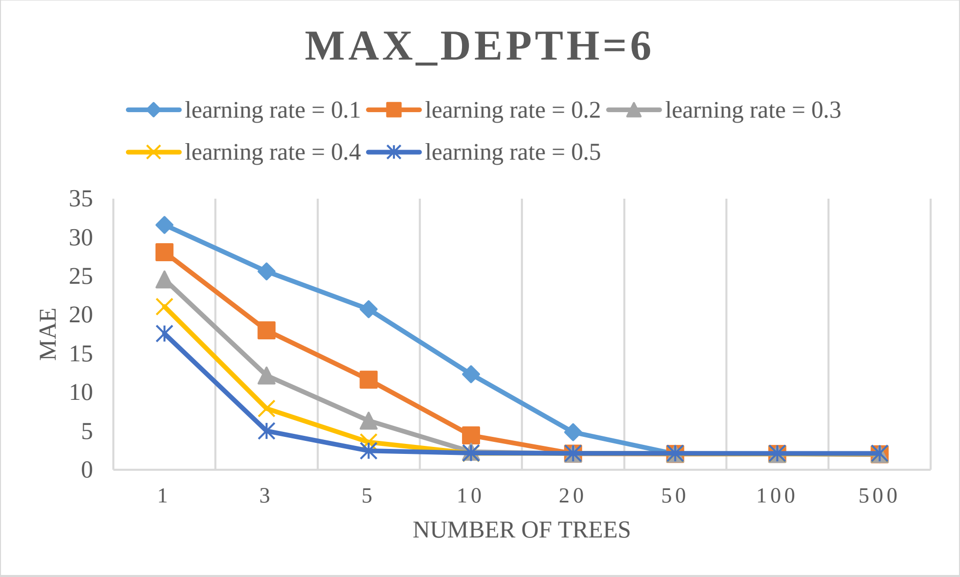


FIGURE 6.2: XGBoost travel time prediction model outputs with the Max_depth=6

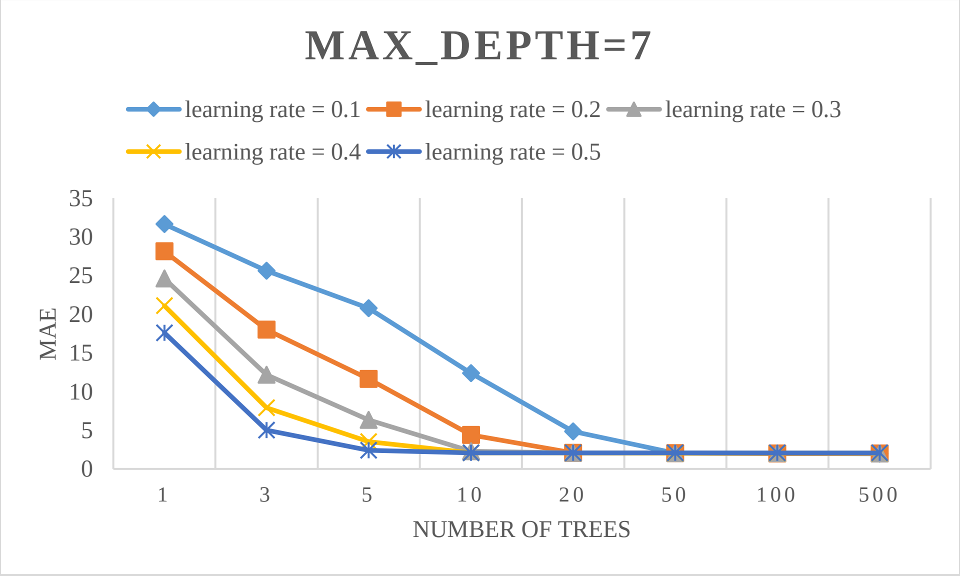


FIGURE 6.3: XGBoost travel time prediction model outputs with the Max_depth=7

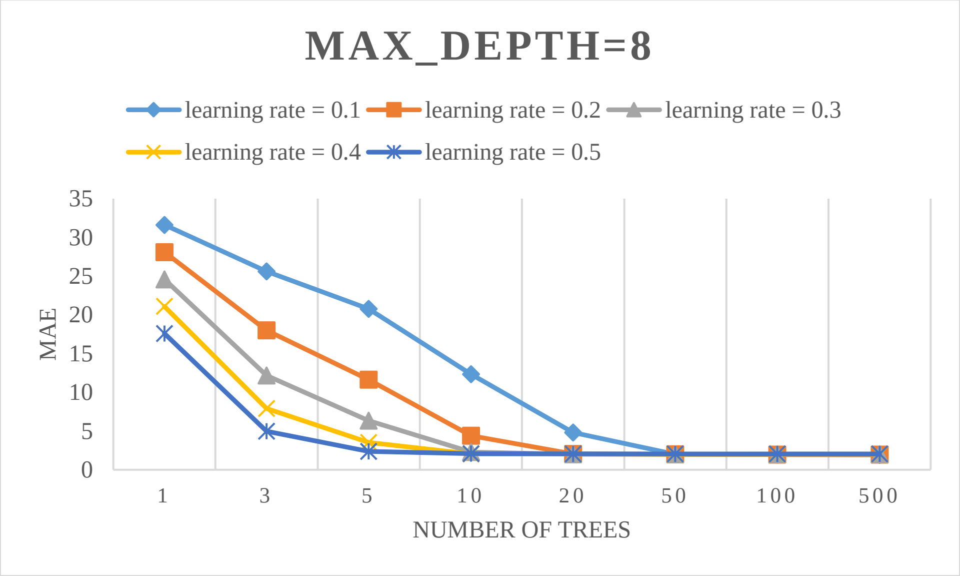


FIGURE 6.4: XGBoost travel time prediction model outputs with the Max_depth=8

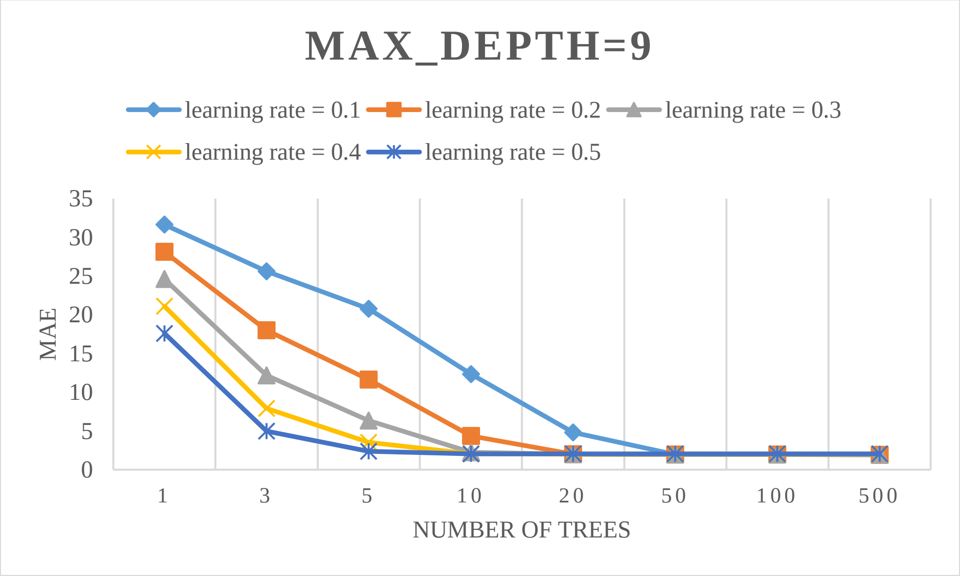


FIGURE 6.5: XGBoost travel time prediction model outputs with the Max_depth=9

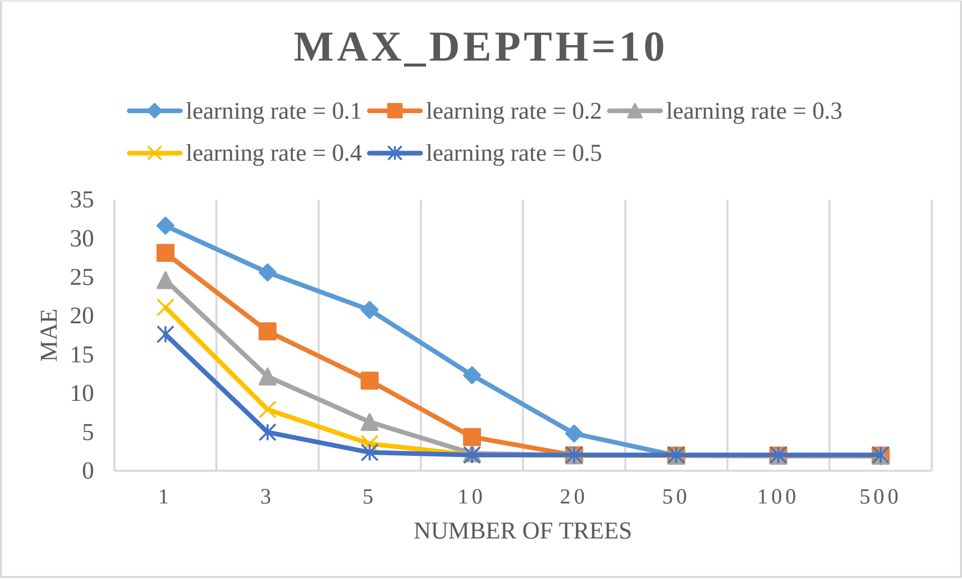


FIGURE 6.6: XGBoost travel time prediction model outputs with the Max_depth=10

Based on Figure 6.1 to Figure 6.6 above, it can be concluded that the MAE value decreases as the number of trees increases, and the slopes of different learning rates are also different. In general, the lower the learning rate is, the higher the initial MAE value (with the number of trees = 1) will be. For example, when the learning rate equals to 0.1, the initial MAE value is about 36.2. In comparison, the figures show that the MAE values are about 17.6 when the number of trees is 1 and the learning rate is 0.5.

Based on the figures above, when the number of trees reaches 50, the value of MAE becomes nearly the same. However, the data in Table 6.3 indicate that the results can still be optimized a little bit if the number of trees keeps increasing. Overfitting is a general problem of traditional ensemble learning methods. For example, the prediction error usually increases when the number of trees increases after it reaches the optimized point in the gradient boosting model (Zhang and Haghani, 2015). In the XGBoost model, the overfitting problem can be solved as the iteration will be stopped when there is no performance improvement after 50 iterations. Therefore, the value 'NA' in Table 6.3 means that the computation already stopped before the number of trees reached those values.

It could be seen that the parameter max_depth does not influence the prediction results significantly since the trends of the errors are nearly the same. However, the data in Table 6.3 shows that as the max_depth increases, the MAE decreases a little bit (the optimized MAEs of max_depth from 5 to 10 are 2.02, 1.98, 1.95, 1.93, 1.91, 1.90, respectively). The data in Table 6.4 shows that as the max_depth increases, the average computation time of the model also decreases a lot, which means the larger value of max_depth can not only increase the accuracy of the model a little bit but also increase the efficiency.

TABLE 6.3: MAEs of different learning rates, number of trees and Max_depth

Learning rate	**MAE**							
Max_depth=5								
	Number of trees							
	1	3	5	10	20	50	100	500
0.1	31.6232	25.6101	20.7441	12.3545	4.86309	2.1066	2.08573	2.01681
0.2	28.105	17.9992	11.6453	4.43742	2.16435	2.11039	2.08105	2.03219
0.3	24.5887	12.1685	6.37425	2.36127	2.14303	2.11073	2.09237	2.05376
0.4	21.0772	7.91567	3.5855	2.19242	2.1688	2.13987	2.11298	NA
0.5	17.5817	5.01915	2.47101	2.22655	2.20399	2.16189	2.13814	NA
Max_depth=6								
	Number of trees							
	1	3	5	10	20	50	100	500
0.1	31.6239	25.6113	20.7459	12.352	4.84463	2.05099	2.03103	1.97875
0.2	28.1064	17.9985	11.6379	4.42059	2.09807	2.05496	2.02066	1.98881
0.3	24.5905	12.1655	6.35543	2.32393	2.09369	2.06605	2.04824	2.02001
0.4	21.0791	7.9191	3.55346	2.12642	2.11301	2.08441	2.06326	NA
0.5	17.5816	5.01106	2.4425	2.16502	2.14012	2.11357	2.11321	NA
Max_depth=7								
	Number of trees							
	1	3	5	10	20	50	100	500
0.1	31.6246	25.6126	20.7487	12.3503	4.83108	2.01236	1.9901	1.95178
0.2	28.1076	18.0042	11.6389	4.40351	2.0696	2.01503	1.997	1.97402

Learning rate	**MAE**							
0.3	24.5923	12.1681	6.33938	2.29773	2.05536	2.02392	2.01867	2.00145
0.4	21.0818	7.90891	3.52957	2.07265	2.06057	2.04777	2.0422	NA
0.5	17.5864	4.99475	2.39171	2.08885	2.07452	2.07155	2.06401	NA
Max_depth=8								
	Number of trees							
	1	3	5	10	20	50	100	500
0.1	31.6262	25.6164	20.7538	12.3515	4.82079	1.98521	1.96533	1.92991
0.2	28.1107	18.0094	11.6371	4.38646	2.04197	1.984	1.968	1.94855
0.3	24.5969	12.1703	6.33313	2.28103	2.01744	1.99892	1.99763	NA
0.4	21.0879	7.90641	3.51711	2.04708	2.02954	2.01933	2.0176	NA
0.5	17.5936	4.9815	2.36822	2.07031	2.05874	2.06077	NA	NA
Max_depth=9								
	Number of trees							
	1	3	5	10	20	50	100	500
0.1	31.6291	25.6203	20.7573	12.3507	4.80693	1.96347	1.94109	1.91498
0.2	28.1168	18.0162	11.6391	4.3746	2.01152	1.95592	1.94509	1.93497
0.3	24.606	12.1708	6.32246	2.25596	1.99238	1.97705	1.97125	NA
0.4	21.0991	7.90241	3.51192	2.02939	2.01599	2.01326	NA	NA
0.5	17.6059	4.97653	2.35518	2.0485	2.03841	2.05002	NA	NA
Max_depth=10								
	Number of trees							
	1	3	5	10	20	50	100	500
0.1	31.6307	25.6248	20.7631	12.352	4.80107	1.94157	1.91908	1.89544
0.2	28.1198	18.0198	11.6369	4.37085	2.00348	1.94777	1.94338	NA
0.3	24.6101	12.1721	6.31815	2.24795	1.98632	1.97365	1.97672	NA
0.4	21.1044	7.89797	3.50731	2.02597	2.01108	2.01454	NA	NA
0.5	17.6118	4.96814	2.3521	2.03704	2.04147	2.05903	NA	NA

TABLE 6.4: Optimized prediction results and computation times

Learning rate	Optimized Result (MAE)	Number of Iterations	Computation Time
Max_depth =5			
0.1	2.01681	500	25 mins
0.2	2.03219	500	25 mins
0.3	2.05376	500	25 mins
0.4	2.079	481	23 mins
0.5	2.11782	217	9 mins
Max_depth =6			
0.1	1.97875	500	25 mins
0.2	1.98881	500	25 mins
0.3	2.02001	500	25 mins
0.4	2.05099	405	20 mins
0.5	2.10818	107	5 mins
Max_depth =7			
0.1	1.95178	500	25 mins
0.2	1.97402	500	25 mins
0.3	2.00145	500	25 mins
0.4	2.03456	231	12 mins
0.5	2.06401	81	4 mins
Max_depth =8			
0.1	1.92991	500	25 mins
0.2	1.94855	500	25 mins
0.3	1.99435	281	17 mins
0.4	2.0176	98	6 mins
0.5	2.05619	73	4 mins
Max_depth =9			
0.1	1.91498	500	25 mins
0.2	1.93497	500	25 mins
0.3	1.96895	167	8 mins
0.4	2.01224	80	4 mins
0.5	2.03841	70	4 mins
Max_depth =10			
0.1	1.89544	500	25 mins
0.2	1.93876	352	18 mins
0.3	1.97233	156	8 mins
0.4	2.00963	74	4 mins
0.8	2.03704	60	4 mins

According to the experimental results, it can be concluded that:

The accuracy level of slower learning rate with a larger number of trees in the model is higher than that of a faster learning rate with a smaller number of trees. The number of trees needed to get optimized result for the model with faster learning rate is also lower than those with slower learning rates.

There is also a need to consider the tradeoff between prediction accuracy and computational time. Since a large number of trees is being fitted, model complexity also increases and requires more computational time. Therefore, the selection of the parameters such as max_depth and number of stopping round is important in the real world.

In addition, the maximum depth of the tree also affects the optimized selection. When the learning rates and number of trees are the same, a higher maximum depth of the tree leads to the lower error rates. A higher max_depth is also more efficient than a lower value since the number of iterations needed to achieve optimized results is lower. In general, a higher max_depth value means a more complex tree model and requires fewer trees to be fitted with a given learning rate.

6.4. Summary

This chapter describes the validation process of the XGBoost-based travel time prediction model. The detailed information about the input features is presented. The parameters of the XGBoost model are also introduced. In order to achieve a better model performance, the parameter tuning process is discussed. The experimental results could give a clear picture of how the analyzed parameters impact the prediction performance.

CHAPTER 7: PREDICTION RESULTS ANALYSIS

7.1. Introduction

This chapter presents the evaluation of the proposed XGBoost model based on the results described in Chapter 6. Section 7.2 presents the analysis of the optimized prediction results from XGBoost model. Section 7.3 presents the performance comparison between the XGBoost model and gradient boosting model. Finally, section 7.4 concludes this chapter with a summary.

7.2. Modelling Results Analysis

In machine learning area, the predictor variables, which are the features mentioned in Chapter 6, usually have significant impacts on the prediction results. Exploring the influence on the individual feature can help understand the data better. Higher relative importance indicates a stronger influence in predicting travel time.

Table 7.1 presents the relative importance of each feature in the optimized XGBoost model. Each predictor variable has a different impact on the predicted travel time. Based on the importance rank of each variable, it can be found that the variable T_{t-1}, which is the travel time at time step t-1 (15 minutes before), contributes the most to the predicted travel time. This result is expected and consistent with a previous study (Zhang and Haghani, 2015), which demonstrates that the immediate previous traffic condition will influence the traffic condition in the future. Therefore, this feature T_{t-1} is the most important and highly correlated with the prediction value.

The results in Table 7.1 show that time of day is the second ranked variable with the relative importance value of 34.85%, and this result is also expected. As mentioned in

Chapter 4, the travel time variability is also highly correlated with the time of day. The travel time usually increases a lot during peak hours and becomes stable during non-peak hours.

The third ranked variable is the segment ID with the relative importance value of 12.65%. The potential reason behind this ranking could be that the segment ID indicates which segment it is. The segment ID contains a lot of potential information such as the geographic location of the segment. Based on the travel time variability analysis results in Chapter 4, different segment locations contribute to different travel time variability characteristics. Therefore, the segment ID is also a necessary and important feature in the model.

Day of week is the 4th ranked variable in the model; the relative importance value of day of week is 3.76%. The variable day of week is also important in the model since the travel time is highly correlated with which day of the week it is. Based on previous studies, the traffic congestion on weekends is less frequent than on weekdays (Chen et al., 2017, Chen et al., 2018). The travel time during peak hours on Friday is usually higher than those on other weekdays (Wang et al., 2017). Therefore, the variable day of week is important in the model; this result is consistent with a previous study (Zhang and Haghani, 2015).

Weather is also considered in the model with a relative importance value of 1.72%. Based on the results mentioned in Chapter 4, inclement weather conditions may have a drastic impact on travel time variability. Therefore, the weather information is also useful in travel time prediction as adverse weather usually increases travel time. This finding is consistent with previous studies (Koesdwiady et al., 2016; Qiao et al., 2016).

The travel time at time step t-1 (15 minutes before) is not the only variable with the consideration of temporal correlation. Several variables such as the travel time of the two steps and three steps ahead (with the relative importance value of 0.40% and 0.33%, respectively) and the travel time change value of the three time step ahead (with the relative importance value of 0.24%, 0.47% and 0.27%, respectively) are considered in the model. These variables are also used in the model of previous studies which had used gradient boosting models to predict freeway travel time (Zhang and Haghani, 2015; Cheng et al., 2018). The time change variables are considered in this study because they could indicate the travel time change trends of the segments. However, the influences of these variables are relatively small. The outcome is similar to the outcome of a previous study (Cheng et al., 2018).

With the consideration of spatial impact, several variables such as the travel time of the two upstream segments (with the relative importance value of 0.29% and 0.40%, respectively) and the travel time of the two downstream segments (with the relative importance value of 0.26% and 0.60%, respectively) one time-step ahead are considered in the model. With respect to the travel time change value, the relative importance values of the two upstream segments are both 0.28%, and the relative importance values of the two upstream segments are 0.36% and 0.69%, respectively. Based on these results, it could be found that the relative importance values of the downstream segments are higher than those of upstream segments. It could be explained by the spatial characteristics of the roadway. If a bottleneck occurs at the downstream segment, the upstream segment will be influenced shortly.

TABLE 7.1: Relative importance of each variable and their ranks in the model

Variable	**Relative Importance (%)**	**Rank**
ID	12.65	3
L	0.24	19
TOD	34.85	2
DOW	3.76	4
Month	2.10	5
Weather	1.72	6
T_{t-1}	38.87	1
T_{t-2}	0.40	10
T_{t-3}	0.33	13
ΔT_{t-1}	0.24	19
ΔT_{t-2}	0.47	9
ΔT_{t-3}	0.27	17
T_{t-1}^{i-1}	0.29	14
T_{t-1}^{i-2}	0.40	10
ΔT_{t-1}^{i-1}	0.28	15
ΔT_{t-1}^{i-2}	0.28	15
T_{t-1}^{i+1}	0.26	18
T_{t-1}^{i+2}	0.60	8
ΔT_{t-1}^{i+1}	0.36	12
ΔT_{t-1}^{i+2}	0.69	7

7.3. Model Comparison

In order to examine the accuracy and effectiveness of the XGBoost model, this section comprehensively evaluates the modeling results of the XGBoost model and compares the results with those of the gradient boosting model. The prediction result of the gradient boosting model is also optimized using a grid search method. For clarity, the mean absolute percentage error (MAPE) is used to evaluate and compare the performance of the two models.

The equation of the MAPE is provided below:

$$MAPE = \frac{100\%}{m}\sum_{i=1}^{m}\left|\frac{y_i - \widehat{y_i}}{y_i}\right|$$

where,

m = The total number of the data.

y_i = The actual travel time value in the test dataset of record i.

$\widehat{y_i}$= The predicted travel time value in the test dataset of record i.

Table 7.2 below presents the comparison between prediction results of the optimized XGBoost model and gradient boosting model. Based on the comparison, it could be concluded that the XGBoost model outperforms the gradient boosting model with both the consideration of accuracy and efficiency. The potential reason behind this could be as follows:

In general, the XGBoost model is a more regularized form of the gradient boosting model. XGBoost uses advanced regularization terms, which improve model generalization capabilities. Therefore, the prediction results of the XGBoost model is more accurate than those of the gradient boosting model. At the same time, the computation time of the XGBoost model (25 mins) is much faster than that of the gradient boosting model (2 hours). One important reason behind the better performance of the XGBoost model could be the parallel processing function. The gradient boosting model is extremely difficult to parallelize since it has sequential characteristics. In comparison, XGBoost can allow us to do the boosting work using distributed processing engines.

Another key reason is the XGBoost model implements the early stopping function, which means that one can stop model assessment when additional trees (see Chapter 6) offer no improvement to the prediction results. This function can help us not only prevent overfitting problem, but also improve the efficiency of the model significantly.

TABLE 7.2: Performance comparison between XGBoost model and gradient boosting model

Number of trees	MAPE XGBoost (%)	MAPE Gradient Boosting (%)
3	14.64	35.10
10	5.22	24.33
20	5.22	16.78
50	4.87	13.56
100	4.82	11.11
200	4.74	9.38
500	4.72	5.67
Average computation time	11.8 mins	Over one hour

7.4. Summary

This chapter presents the numerical results of the developed XGBoost model. The relative importance of each variable in the model is presented and interpreted. In order to examine the accuracy and effectiveness of the proposed model, this chapter also evaluates the optimized modeling results of the proposed XGBoost travel time prediction model and compares them with those of the gradient boosting model. The results demonstrate that the developed XGBoost travel time prediction model significantly improves the computation accuracy and efficiency.

CHAPTER 8: SUMMARY AND CONCLUSIONS

8.1. Introduction

Travel time is an important performance measure for assessing freeway traffic conditions and the extent of highway congestion. Anonymous vehicle probe data is a reliable source for freeway travel time analysis since it greatly improves both data coverage and data fidelity. The travel time variability is highly complex as it is affected by a wide variety of factors. A better understanding of travel time variability patterns can greatly help the decision makers plan, design, operate, and manage a more efficient highway system.

In recent years, with the development of machine learning technologies, various novel algorithms have been developed (Jordan and Mitchell, 2015). One of the representative technologies is the XGBoost model. In recent years, the XGBoost model has gained popularity by winning many data science competitions (e.g., Kaggle competition). Therefore, the XGBoost model has the potential to be applied in transportation-related data analysis fields such as traffic flow, travel speed and travel time prediction.

The primary objective of this research is to develop a methodology for conducting travel time variability analysis for the segments under different traffic conditions. Factors including the time of day, day of week, year, and weather condition are considered in this study. In addition, the XGBoost model-based travel time prediction is presented in this dissertation. The XGBoost prediction model is applied on a real-world freeway corridor so that the gaps between theory and practice can be bridged.

The rest of this chapter is organized as follow: Section 8.2 presents a summary of conclusions of the travel time variability analysis; Section 8.3 presents a summary of conclusions of the numerical results derived from the proposed XGBoost travel time prediction model; Section 8.4 gives a brief discussion of the limitations of the current approaches and provides future research directions.

8.2. Summary and Conclusions of Travel Time Variability Analysis

With the analysis of the travel time variability of eight typical segments on the I-77 southbound corridor in Charlotte, NC, the TTR variability patterns could be identified under different conditions. The information gathered out of the travel time variability analysis in this study can be concluded as follows.

In general, the travel time variability patterns of different segments along the corridor are different. Different cases including PM peak only, AM peak only, double-peak and no peak should be analyzed separately since they demonstrate different results.

With respect to the DOW, the travel time variability analysis results show that for the segments with noticeable peak hour trend, the travel time variability on weekends is much lower than that on weekdays. In particular, for the segments under case 1 and 3 (PM peak only and double peak, respectively), the travel time variability on Friday is the highest. For the segments under case 2 (AM peak only), the travel time variability on Tuesday is the highest. For the segments under case 4 (no peak hour), the travel time variability on each DOW doesn’t change significantly.

With respect to weather conditions, the travel time variability analysis results show that the PTIs under rain condition have obviously higher values than those under normal condition throughout the day. The PTIs under snow/ice/fog condition are also higher than the PTIs under normal condition throughout the day with unique variability patterns.

8.3. Summary and Conclusions of Travel Time Prediction Results

Regarding the travel time prediction, it is found that the XGBoost model can provide reliable prediction results. The relationships between several important parameters in the model (e.g. number of trees, learning rate, and maximum depth of the tree) are discussed in this study. In detail, the accuracy level of a slower learning rate with a larger number of trees in the model is higher than that of a faster learning rate with a smaller number of trees. A higher max_depth value is also more efficient than a lower value since the number of iterations needed to achieve optimized results is smaller.

The relative importance of the features shows that the travel time one step ahead (15 minutes before) contributes the most to the predicted travel time. Features such as the time of day, day of the week and weather also have higher relative importance values in the model than other features.

The proposed XGBoost-based travel time prediction method has considerable advantages over the gradient boosting approach. The performance evaluation result shows that the XGBoost-based model can have better outcomes in terms of both prediction accuracy and efficiency.

8.4. Future Work Directions

The methodology and results for the travel time variability analysis in this study can be helpful for the travel time variability modeling related work in the real world. However, with the limited amount of data, the impacts of accidents and roadworks on travel time variability are not discussed in this study. In the future, the impacts of these variables will be studied if the data can be made available. In detail, the impacts of roadway geometric changes and traffic volume could be further explored. The impact of detailed weather conditions on each roadway segment is also worth exploring particularly when data about more weather stations are available. With respect to different DOW, the potential reason behind the highest travel time variability on Tuesday in Case 2 could be further studied. Furthermore, the travel time variability analysis will be conducted at a network level and relevant characteristics will be examined in detail.

Typically, the XGBoost-based travel time prediction model can provide reliable results with low error rates. However, the impacts of accidents and roadworks on travel time prediction are also worth exploring. In the future, how to incorporate these features in the model will be studied if the data can be made available.

Furthermore, the performance of the travel time prediction model is discussed under all conditions as a whole. In the future, the performances of the model under different traffic conditions (such as both non-congested and congested conditions) can be learned and compared.

REFERENCES

Alajali W. W Zhou, S Wen, and Y Wang. 2018. Intersection traffic prediction using decision tree models. Symmetry 10(9): 386.

Albert LP. 2000. Development and validation of an areawide congestion index using Intelligent Transportation Systems data. Doctoral dissertation, Texas A&M University, College Station, Texas.

Banfield RE, LO Hall, KW Bowyer, and WP Kegelmeyer. 2006. A comparison of decision tree ensemble creation techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(1): 173-180.

Breiman L. 1996. Bagging predictors. Machine learning 24(2): 123-140.

Breiman L. 2001. Random forests. Machine learning 45(1): 5-32.

California. 1998. 1998 California Transportation Plan: Transportation System Performance Measures: Final Report. Transportation System Information Program, California Department of Transportation.

Cambridge Systematics, Inc. 1998. Multimodal corridor and capacity analysis manual. NCHRP Report 399, Transportation Research Board of the National Academies, National Research Council, Washington, D.C.

Cambridge Systematics, Inc., Dowling Associates, Inc., System Metrics Group, Inc., and Texas Transportation Institute. 2008. Cost-effective performance measures for travel time delay, variation, and reliability. NCHRP Report 618, Transportation Research Board of the National Academies, Washington, D.C.

Cambridge Systematics, Inc., High Street Consulting Group, TransTech Management, Inc., Spy Pond Partners, and Ross & Associates. 2009. Performance measurement framework for highway capacity decision making. No. SHRP 2 Report S2-C02-RR, Transportation Research Board of the National Academies, Washington, D.C. http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_rpt_399.pdf.

Charles, E. E. 1997. An introduction to reliability and maintainability engineering. Edition McGraw Hill.

Charlotte C, LM Helenea, and B Sandra. 2017. Empirical estimation of the variability of travel time. Transportation Research Procedia 25: 2769-2783.

Chen C, A Skabardonis, and P Varaiya. 2003. Travel-Time reliability as a measure of service. Transportation Research Record: Journal of the Transportation Research Board 1855:74-79.

Chen P, R Tong, G Lu, and Y Wang. 2018. Exploring travel time distribution and variability patterns using probe vehicle data: case study in Beijing. Journal of Advanced Transportation 3747632.

Chen T, and C Guestrin. 2016. Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, August 13-17.

Chen XM, X Chen, H Zheng, and C Chen. 2017. Understanding network travel time reliability with on-demand ride service data. Frontiers of Engineering Management 4(4): 388-398.

Cheng J, G Li, and X Chen. 2018. Research on travel time prediction model of freeway based on gradient boosting decision tree. IEEE Access 7:7466-7480.

Clark S, and D Watling. 2005. Modelling network travel time reliability under stochastic demand. Transportation Research Part B: Methodological 39(2): 119-140.

Domingos PM. 2012. A few useful things to know about machine learning. Communications of ACM 55(10):78-87.

Dong X, T Lei, S Jin, and Z Hou. 2018. Short-term traffic flow prediction based on xgboost. 2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS). Dali, China, May 24-27.

Dowling RG, A Skabardonis, RA Margiotta, and ME Hallenbeck. 2009. Reliability breakpoints on freeways. Transportation Research Board 88th Annual Meeting, Washington D.C., USA, January 11-15.

Duan Y, Y Lv, and FY Wang. 2016. Travel time prediction with LSTM neural network. 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). Rio de Janeiro, Brazil, November 1-4.

Elefteriadou L, and X Cui. 2007. A framework for defining and estimating travel time reliability. Transportation Research Board 86th Annual Meeting, Washington D.C., USA, January 21-25.

Eliasson J. 2007. The relationship between travel time variability and road congestion. 11th World Conference on Transport Research, Berkeley, California.

Emam E, and H Ai-Deek. 2006. Using real-life dual-loop detector data to develop new methodology for estimating freeway travel time reliability. Transportation Research Record: Journal of the Transportation Research Board 1959: 140-150.
Fan W, and L Gong. 2017. Developing a systematic approach to improving bottleneck analysis in North Carolina. North Carolina DOT. https://connect.ncdot.gov/projects/research/RNAProjDocs/2016-10%20Final%20Report.pdf

Fan SKS, CJ Su, HT Nien, PF Tsai, and CY Cheng. 2018. Using machine learning and big data approaches to predict travel time based on historical and real-time data from Taiwan electronic toll collection. Soft Computing *22*(17):5707-5718.

Florida Department of Transportation. 2011. The Florida reliability method in Florida's mobility performance measures program. http://www.floridampms.com/pdf/trbmpmpaper16.pdf

Freeman EA, GG Moisen, JW Coulston, and BT Wilson. 2015. Random Forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. Canadian Journal of Forest Research 46(3):323-339.

Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. Analysis of Statistics 1189-1232.

Gupta B, S Awasthi, R Gupta, L Ram, P Kumar, BR Prasad, and S Agarwal. 2018. Taxi travel time prediction using ensemble-based random forest and gradient boosting model. In Advances in Big Data and Cloud Computing 63-78.

Han J, J Pei, and M Kamber. 2011. Data mining: concepts and techniques. Elsevier.

Hainen A, J Wasson, S Hubbard, S Remias, G Farnsworth, and D Bullock. 2011. Estimating route choice and travel time reliability with field observations of bluetooth probe vehicles. Transportation Research Record: Journal of the Transportation Research Board 2256: 43-50.

Hamner B. 2010. Predicting travel times with context-dependent random forests by modeling local and aggregate traffic flow. 2010 IEEE International Conference on Data Mining Workshops. Sydney, Australia, December 13.

Hojati AT, L Ferreira, S Washington, P Charles, and A Shobeirinejad. 2016. Modelling the impact of traffic incidents on travel time reliability. Transportation Research Part C: Emerging Technologies 65: 49-60.

Javid RJ, and RJ Javid. 2017. A framework for travel time variability analysis using urban traffic incident data. IATSS Research 42(1):30-38.

Jordan MI and TM Mitchell. 2015. Machine learning: trends, perspectives, and prospects. Science 349(6245):255-260.

Kamga C, and MA Yazıcı. 2014. Temporal and weather-related variation patterns of urban travel time: considerations and caveats for value of travel time, value of variability, and mode choice studies. Transportation Research Part C: Emerging Technologies 45: 4-16.

Kearns M. 1988. Thoughts on hypothesis boosting. Unpublished Manuscript 45:105.

Kim J. 2014. Travel time reliability of traffic networks: characterization, modeling and scenario-based simulation. Doctoral dissertation, Northwestern University, Evanston, Illinois.

Koesdwiady A, R Soua, and F Karray. 2016. Improving traffic flow prediction with weather information in connected cars: a deep learning approach. IEEE Transactions on Vehicular Technology 65(12):9508-9517.

Kwon J, R Barkley, R Hranac, K Petty, and N Compin. 2011. decomposition of travel time reliability into various sources: incidents, weather, work zones, special events, and base capacity. Transportation Research Record: Journal of the Transportation Research Board 2229: 28-33.

LeBlanc M, and R Tibshirani. 1996. Combining estimates in regression and classification. Journal of the American Statistical Association 91(436):1641-1650.

Li X, and R Bai. 2016. Freight vehicle travel time prediction using gradient boosting regression tree. 15th IEEE International Conference on Machine Learning and Applications (ICMLA). Miami, USA, December 9-11.

Li Z, L Elefteriadou, and A Kondyli. 2016. Quantifying weather impacts on traffic operations for implementation into a travel time reliability model. Transportation Letters: The International Journal of Transportation Research 8(1): 47-59.

Liu Y, Y Wang, X Yang, and L Zhang. 2017. Short-term travel time prediction by deep learning: a comparison of different LSTM-DNN models. 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). Yokohama, Japan, October 16-19.

Lomax T, D Schrank, S Turner, and R Margiotta. 2003. Selecting travel time reliability measures. Texas Transportation Institute. https://static.tti.tamu.edu/tti.tamu.edu/documents/TTI-2003-3.pdf.

Lomax T, S Turner, G Shunk, HS Levinson, RH Pratt, PN Bay, and GB Douglas. 1997. Quantifying congestion, Volume 1: Final Report. NCHRP Report 398, Transportation Research Board of the National Academies, Washington, D.C. http://onlinepubs. trb. org/onlinepubs/nchrp/nchrp_rpt_398. pdf.

Martchouk M, F Mannering, and D Bullock. 2010. Analysis of freeway travel time variability using bluetooth detection. Journal of Transportation Engineering 137(10): 697-704.

McLeod DS, L Elefteriadou, and L Jin. 2012. Travel time reliability as a performance measure: applying Florida's predictive model to an entire freeway system. Institute of Transportation Engineers: ITE Journal 82(11): 43.

Min W, and L Wynter. 2011. Real-time road traffic prediction with spatio-temporal correlations. Transportation Research Part C: Emerging Technologies 19(4):606-616.

Moonam HM, X Qin, and J Zhang. 2019. Utilizing data mining techniques to predict expected freeway travel time from experienced travel time. Mathematics and Computers in Simulation 155:154-167.

Myung J, DK Kim, SY Kho, and CH Park. 2011. Travel time prediction using K nearest neighbor method with combined data from vehicle detector system and automatic toll collection system. Transportation Research Record: Journal of the Transportation Research Board 2256:51-59.

Ng M, and ST Waller. 2010. A computationally efficient methodology to characterize travel time reliability using the fast Fourier transform. Transportation Research Part B: Methodological 44(10): 1202-1219.

North Carolina DMV. 2016. North Carolina 2016 Traffic Crash Facts https://connect.ncdot.gov/business/DMV/DMV%20Documents/2016%20Crash%20Facts.pdf

Ogutu JO, HP Piepho, and T Schulz-Streeck. 2011. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proceedings 5(3):11.

Park D and LR Rilett. 1999. Forecasting freeway link travel times with a multilayer feedforward neural network. Computer‐Aided Civil and Infrastructure Engineering 14(5):357-367.

Peer S, CC Koopmans, and ET Verhoef. 2012. Prediction of travel time variability for cost-benefit analysis. Transportation Research Part A: Policy and Practice 46(1): 79-90.

Pu W. 2011. Analytic relationships between travel time reliability measures. Transportation Research Record: Journal of the Transportation Research Board 2254: 122-130.

Qiao W, A Haghani, CF Shao, and J Liu. 2016. Freeway path travel time prediction based on heterogeneous traffic data through nonparametric model. Journal of Intelligent Transportation Systems 20(5):438-448.

Quinlan JR. 1986. Induction of decision trees. Machine learning 1(1):81-106.

Rajsingh EB, J Veerasamy, AH Alavi, and JD Peter. 2018. Advances in Big Data and Cloud Computing (Vol. 645). Springer.

Recker W, Y Chung, J Park, L Wang, A Chen, Z Ji, and JS Oh. 2005, Considering risk-taking behavior in travel time reliability. California Partners for Advanced Transit and Highways (PATH). http://caltrans.ca.gov/newtech/researchreports/2002-2006/2005/to_4110.pdf

Saberi M, and RL Bertini. 2010. Beyond corridor reliability measures: analysis of freeway travel time reliability at the segment level for hot spot identification.

Transportation Research Board 89th Annual Meeting. Washington D.C., USA, January 10-14.

Schrank D, B Eisele, T Lomax, and J Bak. 2015, 2015 Urban Mobility Scorecard. Texas A&M Transportation Institute, INRIX Inc. http://d2dtl5nnlpfr0r.cloudfront.net/tti.tamu.edu/documents/mobility-scorecard-2015.pdf.

Schroeder B, N Rouphail, and S Aghdashi. 2013. Deterministic framework and methodology for evaluating travel time reliability on freeway facilities. Transportation Research Record: Journal of the Transportation Research Board 2396: 61-70.

Shao H, WH Lam, ML Tam, and XM Yuan. 2008. Modelling rain effects on risk-taking behaviors of multi-user classes in road networks with uncertainty. Journal of Advanced Transportation 42(3): 265-290.

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

Strickland, J. (2015). Predictive Analytics Using R. Lulu. com.

Strobl C, AL Boulesteix, A Zeileis, and T Hothorn. 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. Bioinformatics 8(1), p.25.

Texas Transportation Institute, and Cambridge Systems, Inc. (2010), Travel Time Reliability: Making It There on Time, All the Time. FHWA. http://www.ops.fhwa.dot.gov/publications/tt_reliability/TTR_Report.html.

Tsapakis I, T Cheng, and A Bolbol. 2013. Impact of weather conditions on macroscopic urban travel times. Journal of Transport Geography 28: 204-211.

Tu H. 2008. Monitoring travel time reliability on freeways. Doctoral dissertation, Delft University of Technology, Delft, Netherlands.

Tu H, H Li, H Van Lint, V Knoop, and L Sun. 2013. Macroscopic travel time reliability diagrams for freeway networks. Transportation Research Record: Journal of the Transportation Research Board 2396: 19-27.

Turner SM, ME Best, and DL Schrank. 1996. Measures of effectiveness for major investment studies. No. SWUTC/96/467106-1, Transportation Research Board of the National Academies, Washington, D.C. http://tti.tamu.edu/documents/467106-1.pdf

Van Lint J, and H Van Zuylen. 2005. Monitoring and predicting freeway travel time reliability: using width and skew of day-to-day travel time distribution. Transportation Research Record: Journal of the Transportation Research Board 191: 54-62.

Van Lint J, S Hoogendoorn, and H Van Zuylen. 2002. Freeway travel time prediction with state-space neural networks: modeling state-space dynamics with recurrent neural networks. Transportation Research Record: Journal of the Transportation Research Board 1811:30-39.

Vandervalk A, H Louch, J Guerre, and R Margiotta. 2014. Incorporating reliability performance measures into the transportation planning and programming processes: technical reference. No. SHRP 2 Report S2-L05-RR-3, Transportation Research Board of the National Academies, Washington, D.C. https://www.camsys.com/sites/default/files/publications/SHRP2prepubL05Report.pdf.

Wang Z, A Goodchild, and E McCormack. 2017. A methodology for forecasting freeway travel time reliability using GPS data. Transportation Research Procedia 25: 842-852.

Wang D, J Zhang, W Cao, J Li, and Y Zheng. 2018. When will you arrive? estimating travel time based on deep neural networks. Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, USA, February 2-7.

Wang Y, Y Zhang, X Piao, H Liu, and K Zhang. 2018. Traffic data reconstruction via adaptive spatial-temporal correlations. IEEE Transactions on Intelligent Transportation Systems 20(4):1531-1543.

Wang Z, K Fu, and J Ye. 2018. Learning to estimate the travel time. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, United Kingdom, August 19-23.

Wei W, X Jia, Y Liu, and X Yu. 2018. Travel time forecasting with combination of spatial-temporal and time shifting correlation in CNN-LSTM neural network. Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data. Macau, China, July 23-25.

Weng J, L Liu, and J Rong. 2013. Impacts of snowy weather conditions on expressway traffic flow characteristics. Discrete Dynamics in Nature and Society 791743.

Wisitpongphan N, W Jitsakul, and D Jieamumporn. 2012. Travel time prediction using multi-layer feed forward artificial neural network. 2012 Fourth International Conference on Computational Intelligence, Communication Systems and Networks. Phuket, Thailand, July 24-26.

Wolniak M, and S Mahapatra. 2014. Data and performance-based congestion management approach for Maryland highways. Transportation Research Record: Journal of the Transportation Research Board 2420: 23-32.

Wu CH, JM Ho, and DT Lee .2004. Travel-time prediction with support vector regression. IEEE Transactions on Intelligent Transportation Systems 5(4):276-281.

Yang S. 2016. Estimating freeway travel time reliability for traffic operations and planning. Doctoral Dissertation, University of Arizona, Tucson, Arizona.

Yang S. A Malik, and YJ Wu. 2014. Travel time reliability using the Hasofer-Lind-Rackwitz-Fiessler algorithm and kernel density estimation. Transportation Research Record: Journal of the Transportation Research Board 2442: 85-95.

Yazici M, C Kamga, and K Mouskos. 2012. Analysis of travel time reliability in new york city based on day-of-week and time-of-day periods. Transportation Research Record: Journal of the Transportation Research Board 230: 83-95.

Yu B, H Wang, W Shan, and B Yao. 2018. Prediction of bus travel time using random forests based on near neighbors. Computer-Aided Civil and Infrastructure Engineering 33(4):333-350.

Zhang Y, and A Haghani. 2015. A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies 58:308-324.

Zhao L, and SIJ Chien. 2012. Analysis of weather impact on travel speed and travel time reliability. CICTP 2012: Multimodal Transportation Systems—Convenient, Safe, Cost-Effective, Efficient. Beijing, China, August 3-6.

Zheng A, and A Casari. 2018. Feature engineering for machine learning: principles and techniques for data scientists. O'Reilly Media, Inc.

Zheng F, and H Van Zuylen. 2013. Urban link travel time estimation based on sparse probe vehicle data. Transportation Research Part C: Emerging Technologies 31:145-157.

Zheng F, J Li, H Van Zuylen, X Liu, and H Yang. 2017. Urban travel time reliability at different traffic conditions. Journal of Intelligent Transportation Systems 22(2): 106-120.

APPENDIX A: PTIs OF EACH SEGMENT

TMC Code	Year	Time Period	Average PTI	Rating
125N04776	2011	AM Peak	1.06	reliable
		PM Peak	1.13	reliable
	2012	AM Peak	1.04	reliable
		PM Peak	1.40	reliable
	2013	AM Peak	1.04	reliable
		PM Peak	1.60	unreliable
	2014	AM Peak	1.05	reliable
		PM Peak	1.90	unreliable
	2015	AM Peak	1.06	reliable
		PM Peak	2.26	unreliable
125-04776	2011	AM Peak	1.05	reliable
		PM Peak	1.06	reliable
	2012	AM Peak	1.05	reliable
		PM Peak	1.16	reliable
	2013	AM Peak	1.05	reliable
		PM Peak	1.31	reliable
	2014	AM Peak	1.08	reliable
		PM Peak	1.60	unreliable
	2015	AM Peak	1.08	reliable
		PM Peak	1.92	unreliable
125N04777	2011	AM Peak	1.07	reliable
		PM Peak	1.08	reliable
	2012	AM Peak	1.04	reliable
		PM Peak	1.23	reliable
	2013	AM Peak	1.04	reliable
		PM Peak	1.39	reliable
	2014	AM Peak	1.05	reliable
		PM Peak	1.76	unreliable
	2015	AM Peak	1.06	reliable
		PM Peak	2.01	unreliable
125-04777	2011	AM Peak	1.08	reliable
		PM Peak	1.10	reliable
	2012	AM Peak	1.05	reliable
		PM Peak	1.11	reliable
	2013	AM Peak	1.05	reliable
		PM Peak	1.22	reliable
	2014	AM Peak	1.07	reliable

TMC Code	Year	Time Period	Average PTI	Rating
		PM Peak	1.45	reliable
	2015	AM Peak	1.08	reliable
		PM Peak	1.52	unreliable
125N04778	2011	AM Peak	1.09	reliable
		PM Peak	1.16	reliable
	2012	AM Peak	1.07	reliable
		PM Peak	1.16	reliable
	2013	AM Peak	1.06	reliable
		PM Peak	1.24	reliable
	2014	AM Peak	1.08	reliable
		PM Peak	1.44	reliable
	2015	AM Peak	1.09	reliable
		PM Peak	1.58	unreliable
125-04778	2011	AM Peak	1.09	reliable
		PM Peak	1.67	unreliable
	2012	AM Peak	1.07	reliable
		PM Peak	1.65	unreliable
	2013	AM Peak	1.06	reliable
		PM Peak	1.73	unreliable
	2014	AM Peak	1.09	reliable
		PM Peak	1.89	unreliable
	2015	AM Peak	1.10	reliable
		PM Peak	2.04	unreliable
125N04779	2011	AM Peak	1.07	reliable
		PM Peak	1.95	unreliable
	2012	AM Peak	1.05	reliable
		PM Peak	1.92	unreliable
	2013	AM Peak	1.05	reliable
		PM Peak	2.05	unreliable
	2014	AM Peak	1.07	reliable
		PM Peak	2.31	unreliable
	2015	AM Peak	1.10	reliable
		PM Peak	2.61	extremely unreliable
125-04779	2011	AM Peak	1.09	reliable
		PM Peak	2.00	unreliable
	2012	AM Peak	1.07	reliable
		PM Peak	1.98	unreliable
	2013	AM Peak	1.06	reliable
		PM Peak	2.08	unreliable

TMC Code	Year	Time Period	Average PTI	Rating
	2014	AM Peak	1.09	reliable
		PM Peak	2.34	unreliable
	2015	AM Peak	1.11	reliable
		PM Peak	2.70	extremely unreliable
125N04780	2011	AM Peak	1.10	reliable
		PM Peak	2.45	unreliable
	2012	AM Peak	1.07	reliable
		PM Peak	2.43	unreliable
	2013	AM Peak	1.06	reliable
		PM Peak	2.49	unreliable
	2014	AM Peak	1.10	reliable
		PM Peak	2.69	extremely unreliable
	2015	AM Peak	1.12	reliable
		PM Peak	3.19	extremely unreliable
125-04780	2011	AM Peak	1.10	reliable
		PM Peak	2.23	unreliable
	2012	AM Peak	1.08	reliable
		PM Peak	2.18	unreliable
	2013	AM Peak	1.08	reliable
		PM Peak	2.30	unreliable
	2014	AM Peak	1.11	reliable
		PM Peak	2.59	extremely unreliable
	2015	AM Peak	1.13	reliable
		PM Peak	3.20	extremely unreliable
125N04781	2011	AM Peak	1.09	reliable
		PM Peak	2.23	unreliable
	2012	AM Peak	1.07	reliable
		PM Peak	2.20	unreliable
	2013	AM Peak	1.07	reliable
		PM Peak	2.32	unreliable
	2014	AM Peak	1.10	reliable
		PM Peak	2.55	extremely unreliable
	2015	AM Peak	1.13	reliable
		PM Peak	3.24	extremely unreliable

TMC Code	Year	Time Period	Average PTI	Rating
125-04781	2011	AM Peak	1.09	reliable
		PM Peak	1.99	unreliable
	2012	AM Peak	1.08	reliable
		PM Peak	2.02	unreliable
	2013	AM Peak	1.07	reliable
		PM Peak	2.14	unreliable
	2014	AM Peak	1.11	reliable
		PM Peak	2.45	unreliable
	2015	AM Peak	1.14	reliable
		PM Peak	3.24	extremely unreliable
125N04782	2011	AM Peak	1.11	reliable
		PM Peak	2.09	unreliable
	2012	AM Peak	1.10	reliable
		PM Peak	2.18	unreliable
	2013	AM Peak	1.09	reliable
		PM Peak	2.32	unreliable
	2014	AM Peak	1.13	reliable
		PM Peak	2.53	extremely unreliable
	2015	AM Peak	1.17	reliable
		PM Peak	3.27	extremely unreliable
125-04782	2011	AM Peak	1.14	reliable
		PM Peak	1.95	unreliable
	2012	AM Peak	1.15	reliable
		PM Peak	2.01	unreliable
	2013	AM Peak	1.14	reliable
		PM Peak	2.22	unreliable
	2014	AM Peak	1.20	reliable
		PM Peak	2.47	unreliable
	2015	AM Peak	1.31	reliable
		PM Peak	3.22	extremely unreliable
125N04783	2011	AM Peak	1.20	reliable
		PM Peak	2.25	unreliable
	2012	AM Peak	1.22	reliable
		PM Peak	2.27	unreliable
	2013	AM Peak	1.30	reliable

TMC Code	Year	Time Period	Average PTI	Rating
		PM Peak	2.57	extremely unreliable
	2014	AM Peak	1.39	reliable
		PM Peak	2.89	extremely unreliable
	2015	AM Peak	1.62	unreliable
		PM Peak	3.75	extremely unreliable
125-04783	2011	AM Peak	1.62	unreliable
		PM Peak	3.13	extremely unreliable
	2012	AM Peak	1.63	unreliable
		PM Peak	3.03	extremely unreliable
	2013	AM Peak	1.84	unreliable
		PM Peak	3.43	extremely unreliable
	2014	AM Peak	1.94	unreliable
		PM Peak	3.60	extremely unreliable
	2015	AM Peak	2.23	unreliable
		PM Peak	4.71	extremely unreliable
125N04784	2011	AM Peak	1.60	unreliable
		PM Peak	3.10	extremely unreliable
	2012	AM Peak	1.75	unreliable
		PM Peak	3.34	extremely unreliable
	2013	AM Peak	2.01	unreliable
		PM Peak	4.04	extremely unreliable
	2014	AM Peak	2.33	unreliable
		PM Peak	4.09	extremely unreliable
	2015	AM Peak	2.77	extremely unreliable
		PM Peak	5.45	extremely unreliable
125-04784	2011	AM Peak	1.37	reliable
		PM Peak	1.55	unreliable
	2012	AM Peak	1.56	unreliable

TMC Code	Year	Time Period	Average PTI	Rating
		PM Peak	2.04	unreliable
	2013	AM Peak	1.86	unreliable
		PM Peak	2.89	extremely unreliable
	2014	AM Peak	2.12	unreliable
		PM Peak	2.88	extremely unreliable
	2015	AM Peak	2.65	extremely unreliable
		PM Peak	3.75	extremely unreliable
125N04785	2011	AM Peak	1.38	reliable
		PM Peak	1.71	unreliable
	2012	AM Peak	1.60	unreliable
		PM Peak	2.08	unreliable
	2013	AM Peak	1.87	unreliable
		PM Peak	2.95	extremely unreliable
	2014	AM Peak	2.11	unreliable
		PM Peak	2.85	extremely unreliable
	2015	AM Peak	2.63	extremely unreliable
		PM Peak	3.61	extremely unreliable
125-04785	2011	AM Peak	1.29	reliable
		PM Peak	1.44	reliable
	2012	AM Peak	1.49	reliable
		PM Peak	1.66	unreliable
	2013	AM Peak	1.73	unreliable
		PM Peak	2.13	unreliable
	2014	AM Peak	2.00	unreliable
		PM Peak	2.23	unreliable
	2015	AM Peak	2.47	unreliable
		PM Peak	2.79	extremely unreliable
125N04786	2011	AM Peak	1.19	reliable
		PM Peak	1.16	reliable
	2012	AM Peak	1.34	reliable
		PM Peak	1.21	reliable
	2013	AM Peak	1.50	reliable

TMC Code	Year	Time Period	Average PTI	Rating
		PM Peak	1.37	reliable
	2014	AM Peak	1.78	unreliable
		PM Peak	1.46	reliable
	2015	AM Peak	2.22	unreliable
		PM Peak	1.54	unreliable
125-04786	2011	AM Peak	1.13	reliable
		PM Peak	1.10	reliable
	2012	AM Peak	1.22	reliable
		PM Peak	1.08	reliable
	2013	AM Peak	1.42	reliable
		PM Peak	1.09	reliable
	2014	AM Peak	1.66	unreliable
		PM Peak	1.14	reliable
	2015	AM Peak	2.13	unreliable
		PM Peak	1.12	reliable
125N04787	2011	AM Peak	1.13	reliable
		PM Peak	1.09	reliable
	2012	AM Peak	1.20	reliable
		PM Peak	1.08	reliable
	2013	AM Peak	1.35	reliable
		PM Peak	1.09	reliable
	2014	AM Peak	1.52	unreliable
		PM Peak	1.13	reliable
	2015	AM Peak	1.91	unreliable
		PM Peak	1.15	reliable
125-04787	2011	AM Peak	1.61	unreliable
		PM Peak	1.12	reliable
	2012	AM Peak	1.66	unreliable
		PM Peak	1.18	reliable
	2013	AM Peak	1.77	unreliable
		PM Peak	1.24	reliable
	2014	AM Peak	1.93	unreliable
		PM Peak	1.46	reliable
	2015	AM Peak	2.20	unreliable
		PM Peak	1.50	unreliable
125N04788	2011	AM Peak	1.81	unreliable
		PM Peak	1.08	reliable
	2012	AM Peak	1.90	unreliable
		PM Peak	1.07	reliable
	2013	AM Peak	2.09	unreliable

TMC Code	Year	Time Period	Average PTI	Rating
		PM Peak	1.11	reliable
	2014	AM Peak	2.32	unreliable
		PM Peak	1.26	reliable
	2015	AM Peak	2.62	extremely unreliable
		PM Peak	1.25	reliable
125-04788	2011	AM Peak	1.72	unreliable
		PM Peak	1.09	reliable
	2012	AM Peak	1.71	unreliable
		PM Peak	1.06	reliable
	2013	AM Peak	2.09	unreliable
		PM Peak	1.07	reliable
	2014	AM Peak	2.14	unreliable
		PM Peak	1.11	reliable
	2015	AM Peak	2.63	extremely unreliable
		PM Peak	1.13	reliable
125N04789	2011	AM Peak	1.41	reliable
		PM Peak	1.13	reliable
	2012	AM Peak	1.42	reliable
		PM Peak	1.10	reliable
	2013	AM Peak	1.61	unreliable
		PM Peak	1.10	reliable
	2014	AM Peak	1.67	unreliable
		PM Peak	1.12	reliable
	2015	AM Peak	1.97	unreliable
		PM Peak	1.11	reliable
125-04789	2011	AM Peak	1.08	reliable
		PM Peak	1.06	reliable
	2012	AM Peak	1.09	reliable
		PM Peak	1.04	reliable
	2013	AM Peak	1.19	reliable
		PM Peak	1.04	reliable
	2014	AM Peak	1.23	reliable
		PM Peak	1.05	reliable
	2015	AM Peak	1.42	reliable
		PM Peak	1.06	reliable
125N04790	2011	AM Peak	1.06	reliable
		PM Peak	1.06	reliable
	2012	AM Peak	1.05	reliable

TMC Code	Year	Time Period	Average PTI	Rating
		PM Peak	1.04	reliable
	2013	AM Peak	1.07	reliable
		PM Peak	1.05	reliable
	2014	AM Peak	1.15	reliable
		PM Peak	1.06	reliable
	2015	AM Peak	1.24	reliable
		PM Peak	1.06	reliable
125-04790	2011	AM Peak	1.05	reliable
		PM Peak	1.05	reliable
	2012	AM Peak	1.04	reliable
		PM Peak	1.04	reliable
	2013	AM Peak	1.04	reliable
		PM Peak	1.04	reliable
	2014	AM Peak	1.06	reliable
		PM Peak	1.05	reliable
	2015	AM Peak	1.07	reliable
		PM Peak	1.05	reliable
125N04791	2011	AM Peak	1.07	reliable
		PM Peak	1.07	reliable
	2012	AM Peak	1.06	reliable
		PM Peak	1.05	reliable
	2013	AM Peak	1.05	reliable
		PM Peak	1.05	reliable
	2014	AM Peak	1.07	reliable
		PM Peak	1.06	reliable
	2015	AM Peak	1.07	reliable
		PM Peak	1.07	reliable
125-04791	2011	AM Peak	1.23	reliable
		PM Peak	1.25	reliable
	2012	AM Peak	1.22	reliable
		PM Peak	1.24	reliable
	2013	AM Peak	1.23	reliable
		PM Peak	1.28	reliable
	2014	AM Peak	1.16	reliable
		PM Peak	1.08	reliable
	2015	AM Peak	1.06	reliable
		PM Peak	1.07	reliable

APPENDIX B: SEGMENT LENGTH INFORMATION

TMC Code	Segment Length (miles)
125-04791	0.54
125N04791	0.62
125-04790	2.25
125N04790	0.56
125-04789	1.65
125N04789	0.98
125-04788	0.11
125N04788	0.53
125-04787	0.49
125N04787	0.95
125-04786	0.02
125N04786	0.80
125-04785	0.09
125N04785	0.20
125-04784	0.04
125N04784	0.94
125-04783	0.58
125N04783	0.22
125-04782	0.62
125N04782	0.38
125-04781	0.74
125N04781	0.22
125-04780	0.16
125N04780	0.26
125-04779	0.67
125N04779	0.56
125-04778	0.57
125N04778	0.42
125-04777	0.46
125N04777	0.82
125-04776	0.01
125N04776	1.67