Center for Advanced Multimodal Mobility Solutions and Education

Project ID: 2020 Project 01

Travel Time Forecasting on a Freeway Corridor: a Dynamic Information Fusion Model based on the Random Forests Approach

Final Report

by

Wei Fan (ORCID ID: https://orcid.org/0000-0001-9815-710X)
Bo Qiu (ORCID ID: https://orcid.org/0000-0001-7096-6837)

Wei Fan, Ph.D., P.E.
Director, USDOT CAMMSE University Transportation Center
Professor, Department of Civil and Environmental Engineering
The University of North Carolina at Charlotte
EPIC Building, Room 3261, 9201 University City Blvd, Charlotte, NC 28223
Phone: 1-704-687-1222; Email: wfan7@uncc.edu

for

Center for Advanced Multimodal Mobility Solutions and Education
(CAMMSE @ UNC Charlotte)
The University of North Carolina at Charlotte
9201 University City Blvd
Charlotte, NC 28223

September 2021

ACKNOWLEDGEMENTS

This project was funded by the Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE @ UNC Charlotte), one of the Tier I University Transportation Centers that were selected in this nationwide competition, by the Office of the Assistant Secretary for Research and Technology (OST-R), U.S. Department of Transportation (US DOT), under the FAST Act. The authors are also very grateful for all of the time and effort spent by DOT and industry professionals to provide project information that was critical for the successful completion of this study.

DISCLAIMER

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

TMS	Traffic Management Systems
SVR	Support Vector Regression
RF	Random Forest
XGBoost	eXtreme Gradient Boosting
K-NN	K-Nearest Neighbor
LSTM	Long Short-Term Memory
RITIS	Regional Integrated Transportation Information System
TOD	Time of Day
DOW	Day of Week
TTP	Travel Time Prediction
MAPE	Mean Absolute Percentage Error
RMSE	Root Mean Square Error
MRE	Mean Relative Error

EXECUTIVE SUMMARY

Metropolitan areas suffer from frequent road traffic congestion not only during peak hours but also during off-peak periods. Different machine learning methods have been employed in travel time prediction; however, such machine learning methods practically face the problem of overfitting. Tree based ensembles have been applied in various prediction field, and such approaches usually produce high prediction accuracy by aggregating and averaging individual decision trees. The inherent advantages of these approaches are that they can get better prediction results while also having a good bias-variance trade off which can help to avoid the overfitting. However, the reality is that the application of tree-based integration algorithms in traffic prediction is still limited. In order to improve the accuracy and the interpretability of the model, random forest (RF) is used to analyze and model the travel time on freeways.

As the traffic conditions often greatly change, the prediction results are often unsatisfactory. In order to improve the accuracy of short-term travel time prediction in the freeway network, a practically feasible and computationally efficient RF prediction method for real-world freeways by using probe traffic data was generated. In addition, the variables' relative importance can also be ranked, which provides an investigation platform to gain a better understanding of how different contributing factors might affect travel time on freeways. This research develops an RF method to predict the freeway travel time by using the probe vehicle-based traffic data and weather data. Detailed information about the input variables and data pre-processing are presented. To measure the effectiveness of proposed travel time prediction algorithms, the mean absolute percentage errors (MAPE) are computed for different observation segments combined with different prediction horizon ranging from 15 to 60 minutes.

The parameters of the RF model are estimated by using the training sample set. After the parameter tuning process is completed, the proposed RF model is developed. The features' relative importance showes that the variables travel time 15 minutes before and time of day (TOD) contribute the most to the predicted travel time result. The model performance is also evaluated and compared against extreme gradient boosting method, and the results indicates that the RF always produces more accurate travel time prediction.

CHAPTER 1: INTRODUCTION

1.1. Problem Statement and Motivation

Nowadays, travel time prediction plays a significant role as it can greatly help route planning and also the development of countermeasures to reduce traffic congestion. Metropolitan areas are adversely affected by frequent road traffic congestion not only in peak hours but also in off-peak periods. Therefore, the capability to forecast traffic conditions, particularly travel times, is of utmost importance in traffic management applications aimed at relieving negative social, environmental and economic impacts for people. The definition of travel time is the total time for a vehicle to travel from one point to another over a specified route (Zhu et al., 2009). Recently, the need for travel time prediction has become indispensable due to the increasing congestion in the roadway network. However, travel time prediction is highly complex as it is affected by a wide variety of factors. Metropolitan areas are suffering frequent road traffic congestion not only in peak hours but also in off-peak time periods. Accurate and reliable travel time prediction in freeway networks is a critical component that will be helpful to all modes of transportation in all urban, suburban and rural areas. It is widely accepted that considerable accuracy and reliability of travel time prediction is highly desired for both travelers and transportation planners. Therefore, the capability to forecast dynamically changing traffic conditions, particularly travel times, is of utmost importance in a wide range of traffic management applications aimed at relieving its negative impact on society, environment and economy. Accurate travel time prediction can greatly help enhance the performance of the traffic management systems (TMS), in which travelers are given the opportunities to react to the traffic proactively (Oh et.al., 2015). The acquisition and popularization of big data in the field of transportation have enabled the collection and diffusion of real-time traffic information. Different machine learning approaches

have been employed by different researchers, and the results indicated that such approaches can give better performances than traditional models. However, such machine learning methods are practically faced with an overfitting problem that is difficult to overcome. Especially, when the traffic conditions greatly change, the prediction results are often unsatisfactory. In addition, the RF method has a very good Bias-Variance trade-off which can help avoid the overfitting problem. This research develops an RF method to predict the freeway travel time by using the probe vehicle-based traffic data, and therefore helps to gain a better understanding of how different contributing factors might affect travel time on freeways. In this study, as the second ensemble tree based machine learning method, eXtreme Gradient Boosting (XGBoost) is also deployed, which is an algorithm that has ensemble of decision trees and is robust to outliers. XGBoost algorithm is believed to have a good performance on time series predictions (Kankanamge et al., 2019). To validate the effectiveness of two travel time prediction methods, the proposed approaches are tested using a freeway corridor in Charlotte, North Carolina using the probe vehicle-based traffic data. A comparison will be made, the proposed method will achieve a better prediction performance not only in accuracy but also in stability.

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

1. To select the most appropriate travel time prediction variables that could be used to accurately predict results;
2. To systematically analyze the travel time with the consideration of time of day, day of week, month and weather. The potential significant impact factors are analyzed and ranked;

3. To select a real-world freeway corridor to examine the developed prediction models so that the gaps between the theoretical research and the application of the developed travel time prediction model can be bridged.

1.2. Objectives

This research is intended to develop and compare advanced machine learning-based approaches (e.g., RF and XGBoost method) are employed to predict the freeway travel time. The expected contributions from this research are summarized as follows:

1. Ability to select appropriate machine learning methods to predict travel time and identify the most impact traffic variables;
2. Ability to understand the travel time of selected segments with the consideration of time of day, day of week, month and weather.

1.3. Report Overview

The research will be structured as follows.

In Chapter 1, the background and motivation of the travel time prediction has been discussed, followed by the description of study objectives and expected contributions.

Chapter 2 presents a comprehensive review of the current state-of-the-art and state-of-the-practice on short-term travel time prediction. Research on more reliable short-term travel time forecasting has attracted numerous researchers from transportation engineers to data scientists in the last several decades. The machine learning (data-based parametric models and non-parametric models) traffic prediction methods will be introduced.

Chapter 3 presents the RITIS data set that is used to analyze travel time prediction, including the travel time data and historical weather data utilized in this study. The detailed information about

the raw travel time data sources are described first, followed by the discussions about weather data collection. The data processing steps are also described in detail in this chapter.

Chapter 4 discusses the travel time prediction methodology based on the data described in Chapter 3. The Random Forests based travel time prediction model will be developed. The detailed process will be described such as the data structure configuration, parameter determination, model training, and model validation.

Chapter 5 presents the validation of the proposed machine learning models based on the data described in Chapter 3. For the machine learning prediction model, the data training step will be described to determine the parameters in the model structure. The potential parameters will include but are not limited to: time of day, day of week, month of year, weather conditions, segment characteristics, etc. The evaluation of the proposed machine learning models based on the data described in previous Chapters. The MAPE will be used to measure the prediction error. Potential impacts of the introduction of new methodology on the efficiency will be discussed.

Chapter 6 concludes the report with a summary of the developed prediction models, solution approaches, and research results. Suggestions for future research will be also provided.

CHAPTER 2: LITERATURE REVIEW

2.1. Introduction

This chapter provides a comprehensive review of various aspects related to travel time studies and travel time prediction methodologies. This should give a clear picture of existing current efforts toward the modeling of travel time prediction.

The following sections are organized as follows. Section 2.2 presents the definitions of travel time and classification methods. Section 2.3 gives a comprehensive review of existing methods of travel time prediction, which include statistical methods and machine learning methods. Furthermore, the section shows the common research and methods that have been applied to predict travel time. Finally, section 2.4 concludes this chapter with a summary.

2.2. Travel Time Prediction

2.2.1. Definitions of Travel Time

Travel time is defined as the total time for a vehicle to travel from one point to another over a specified route (Zhu et al., 2009).

2.2.2. Classification Approach

Transportation researchers and data scientists have developed various techniques in the past three decades to provide more reliable future travel time estimation methods (Oh et al., 2015). Generally speaking, such techniques can be classified into three groups: naive methods, traffic theory-based methods and data-driven methods. As the name indicates, the naive prediction models are very simple methods, which typically do not involve the estimation of model parameters. As the model assumptions are usually restrictive, they are not actually fulfilled in many situations (Wunderlich et al., 2000). As one of the traffic theorybased methods, traffic flow simulation and user-optimal dynamic traffic assignment have been widely used in

freeway travel time prediction. Examples include Papageorgiou et al. (2010) and Dion et al. (2004). In data-based traffic time prediction models, the function that relates traffic factors with the prediction result (dependent variable) is not obtained from predetermined traffic theory, as the relationships of variables come from the sample data itself by using statistical data mining methods. This approach greatly expands the pool of researchers who can participate in travel time prediction because they no longer have to become experts in traffic theory. However, such data-based methods usually need a lot of data, which is not always available. The data-based models are strongly subjected to data availability and accessibility (Van Lint, 2006).

Travel time prediction can be categorized from different perspectives, and the most popular classification method is to categorize them according to its prognosis horizon as short, medium, and long-term (Oh et al., 2018). Van Lint (2004) defines the short-term travel time prediction at 0-60 minutes intervals. It was found that making the appropriate time horizon in travel time prediction plays the most significant role in the travel time prediction applications (Shen, 2008). And the second important perspective is the road network category, including either arterial roads or freeways. It is more complicated to make prediction on the urban signalized arterial roads due to the signal cycles and intersections (Oh et al., 2018). In short term traffic flow prediction, researchers consider the flows, speeds, densities and travel time, which are important components of the application of ITS (Liu et al., 2017).

2.3. Travel Time Prediction Approach

Thanks to the integration of big data and transportation management, different kinds of approaches have been studied and applied in this area. The approaches can be divided into two general groups: statistical methods and machine learning methods. In statistical models, such as the linear regression and time series, the autoregressive integrated moving average (ARIMA)

model has been widely applied to predict travel times based on the historical data. Machine learning methods are considered more effective, accurate and feasible. Different machine approaches (such as neural network, ensemble learning and support vector machines) have been employed by different researchers, and the results indicate that such approaches can give better performances than traditional statistical models (Mori et al., 2015).

Research on more reliable short-term travel time forecasting has attracted numerous researchers from transportation engineers to data scientists in the last several decades. The machine learning (data-based) traffic prediction methods can be divided into two major categories: parametric models and non-parametric models (Van, 2004). Parametric models are always model-based methods, where all of the parameters can be estimated with empirical data and the model structure is predetermined based on certain theoretical assumptions. Linear regression is the most typical parametric model, where the dependent variable is a linear function of the explanatory (independent) input variables. The input variables are typically traffic observations in several past time intervals. Bayesian net is the second type of the parametric models, where the explanatory variables are assumed to be conditionally independent, given the target variable. The third type of the parametric models is time series models, which is a series of data points indexed in time order. Time series forecasting involves the use of a model to predict future values based on previously observed values. Autoregressive integrated moving average (ARIMA) model is the most widely used for travel time prediction. The first application of ARIMA in the field of traffic analysis dates back to 1979 (Ahmed & Cook, 1979). For parameter-based approaches, the integrity of real-time data is also a critical factor that determines the prediction accuracy, since many model-based systems deal with feeding data in real-time for online services.

In the non-parametric models, both the structure of the model and the parameters are not predetermined. However, the term “non-parametric” does not mean that there are no parameters in the models to be estimated. Furthermore, the number and typology of the parameters are unknown a priori and sometimes uncountable. Due to the rapid development of data science, non-parametric estimation methodologies are being quickly updated. The non-parametric models tend to be more efficient and therefore the more advanced model structure. It is mentioned that the efficiency of data-driven approaches, in general, is poor and not fit for real-time applications (Oh et al., 2015). One of the most popular in the literature of travel time prediction is the artificial neural networks (ANN). Due to their ability to capture complex relationships in large data sets, ANN methods have been widely used in travel time forecasting (Dharia & Adeli, 2003). As the typical non-parametric models, ANN can be developed without being given a specific form of the function. Furthermore, the restrictions on the multicollinearity of the explanatory variables can be partially overcome. Different types of neural networks have been applied to travel time forecasting, such as the regular multilayer feedforward neural networks (Yildirimoglu & Geroliminis, 2013) and spectral basis neural networks (Park & Rilett, 1999). The input variables selection is different which depends on the data availability and the model training process, and different types of neural networks can be carried out by different variations of the backward algorithm. Support vector machine (SVM) methods are another choice for travel time prediction. This advanced algorithm consists of decision function, the application of the kernel functions and the sparsity of solutions. The SVM model is good for travel time prediction based on historical travel time data. Some researchers (Yildirimoglu & Geroliminis, 2013; Wu et al, 2004) used SVM methods to estimate travel time. In the calculation process, the algorithm maps the input data into a higher dimensional space by the kernel function. The process stops

after finding the flattest linear function which relates to the transferred input vectors (i.e., when the target variable with an error smaller than a predefined threshold). This linear function can be mapped again into the initial space and get the final nonlinear function which is used for travel time prediction. Both the ANN and SVM models tend to be overfitting due to the complicated structure and the large number of parameters that need to be calibrated, which is a serious problem commonly existed in the non-parameter machine learning algorithm.

Another popular non-parametric approach in the travel time prediction applications is the local regression approach. Local linear regression can be used to optimally balance the use of historical and real time data (Rupnik et al., 2015), which is able to yield accurate prediction results. In the local regression algorithm, a set of historical data which have similar characteristics to the current situation will be selected by the algorithm. The prediction results base themselves on generating a model constructed by the chosen data set. The types of the local regression models depend on the techniques used to select the set of similar historical points and also depend on the methodology chosen to fit the model.

Some semi-parametric models have been developed in traffic time prediction, which are a combination of parametric and non-parametric methods. The main idea of the semi-parametric method is to loosen some of the assumptions of the parametric model to obtain a more flexible structure (Ruppert et al., 2003). In the case of application, semi-parametric models are presented in the form of varying coefficient regression models. Travel time can be calculated by a linear function of the naive historical and instantaneous predictors. Furthermore, the parameters vary depending on the departure time interval and prediction horizon (Schmitt & Jula, 2007).

With the wide applications of machine learning algorithm in the field of travel time prediction, different approaches have been deployed in different area with different types of data source.

The methodologies have been used by researchers include, but are not limited to, the following: SVM, neural network (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., RF and gradient boosting), etc. Some research details are listed as following and Table 2.1 provides a summary of the studies reviewed in chronological order.

2.3.1. Support Vector Regression Approach

2.3.1.1. Wu et al.'s research work

Support Vector Regression (SVR) was applied by Wu et al. (2004) for travel-time prediction and its results were compared to other travel time prediction methods as baseline using real highway traffic data. As SVR has greater generalization ability and guarantee global minima for given training data, it was believed that SVR has a better performance than time series method. The results showed that the SVR predictor 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.

2.3.2. Nearest Neighbors Approach

2.3.2.1. Myung et al.'s research work

Myung et al. (2011) proposed a model to predict travel times on the basis of the k nearest neighbor (KNN) method using data provided by the vehicle detector system and the automatic toll collection system. By combining these two sets of data, the model minimized the limitations of each dataset and enhanced the prediction's accuracy. The authors compared the prediction results of the proposed model with the predictions of other models by using actual data. The comparison results showed that the proposed model predicts travel times much more accurately.

2.3.2.2. Yu et al.'s research work

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

2.3.2.3. Moonam et al.'s research work

Moonam et al. (2019) conducted a study to predict the expected travel time based on the experienced travel time using the data mining techniques such as k-nearest neighbor (k-NN), least squares regression boosting (LSBoost) and Kalman filter (KF) methods. The authors compared the performances of each methods from both link and corridor perspectives and concluded that the KF method offers superior prediction accuracy in a link-based model.

2.3.3. Neural Network Approach

2.3.3.1. 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 collected on freeway of Houston, Texas, by the automatic vehicle identification (AVI) system was used as the validation database. The proposed model could provide acceptable prediction results with the MAPE range being from 7.4% to 18%.

2.3.3.2. Van Lint et al.'s research work

Van lint et al. (2002) presented an approach to predict freeway travel time based on state-space neural network. The data from freeway operations simulation (FOSIM) 4.1 was 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 loss of predictive performance.

2.3.3.3. Wisitpongphan et al.'s research work

Wisitpongphan et al. (2012) proposed a BP neural network model to predict freeway link travel time. The one-month vehicle trajectories data of 297 probes 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) less than 3%.

2.3.3.4. 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 proposed 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 an analytical estimation model. The performance of these two models was evaluated with data derived from VISSIM simulation model. The final results suggested that the Artificial Neural Network model performs better.

2.3.3.5. Duan et al.'s research work

Duan et al. (2016) employed a LSTM neural network model to predict freeway travel time. The authors constructed 66 series LSTM neural networks by using travel time data 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 result.

2.3.3.6. 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 highway in California, US. The results of proposed model were compared with the results of other regression models and ARIMA model. The comparison results showed that the performance of the LSTM neural network model was the

best.

2.3.3.7. Wang et al.'s research work

Wang et al. (2018) presented a novel machine learning method to predict the vehicle travel time based on 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 our proposed deep learning algorithm significantly outperforms the state-of-the-art algorithms.

2.3.4. Ensemble Learning Approach

2.3.4.1. Hamner et al.'s research work

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

2.3.4.2. 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.

2.3.4.3. Li and Bai's research work

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

2.3.4.4. Gupta et al.'s research work

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

Table 2.1 Summary of Travel Time Prediction Using Machine Learning Approaches

Year	Author	Country/City	Roadway Category	Data Source	Method Category	Data Type	Prediction method
2000	Wunderlich et al.	N/A	N/A	Simulated data from INTEGRATION	Navie model	Travel time	Exponential filtering
2002	Dion et al.	Virginia, US	N/A	Simulated data from INTEGRATION	Traffic theory-base model	Travel time	Delay models
2002	Van Lint et al.	N/A	Freeway	Simulated data from FOSIM	Non-parametric	Travel time, travel speed	State-Space Neural Network
2005	Wu et al.	Taiwan	Highway	Loop detector	Non-parametric	Travel speed	SVR
2007	Schmitt and Jula	California, US	Urban road	Loop detector	Navie model	Travel time	Switch model
2008	Zou et al.	Maryland, US	Highway	Roadside detector	Hybird non-parametric	Travel time	Combined Clustering Neural Networks
2009	Li et al.	Atlanta, US	N/A	simulated data from VISSIM	Hybird non-parametric	Travel time, travel speed	Combined Boosting and Neural Network
2010	Papageorgiou et al.	N/A	N/A	simulated data from MATANET	Traffic theory-base model	Travel time	Macroscopic Simulation
2010	Hamner et al.	N/A	N/A	GPS	Non-parametric	Travel speed	RF
2011	Myung et al.	Korea	N/A	ATC system	Non-parametric	Travel time	KNN
2012	Wisitpongphan	Bangkok, Thailand	Highway	GPS	Non-parametric	Travel time	BP Neural Network
2013	Yildirimoglu & Geroliminis's	California, US	Freeway	Loop detector	Hybird non-parametric	Travel time	Combined Gaussian Mixture, PCA, and Clustering
2015	Zhang and Haghani	Maryland, US	Interstate highway	INRIX	Non-parametric	Travel time	Gradient boosting
2015	Joao et al.	Porto, Portugal	Urban road	STCP system	Hybird non-parametric	Travel time	Combined RF, Projection Pursuit Regression and SVM
2016	Duan et al.	England	Highway	Cameras, GPS and loop detectors	Non-parametric	Travel time	LSTM Neural Network
2016	Li and Bai	Ningbo, China	N/A	N/A	Non-parametric	Truck trajectory, travel time, travel speed	Gradient boosting
2017	Liu et al.	California, US	Interstate highway	PeMS	Non-parametric	Travel time	LSTM Neural Network

Year	Author	Country/City	Roadway Category	Data Source	Method Category	Data Type	Prediction method
2017	Fan et al.	Taiwan	Highway	Electric toll	Non-parametric	Travel time, vehicle information	RF method
2017	Yu et al.	Shenyang, China	bus route	AVL system	Non-parametric	Bus travel time	RF and KNN
2018	Wang et al.	Beijing, China	Urban road	Floating Car Data	Non-parametric	Taxi ravel time, vehicle trajectory data	LSTM Neural Network
2018	Wei et al.	China	Urban road	Vehicle passage records	Non-parametric	Travel time	LSTM Neural Network
2018	Wang et al.	Beijing and Chengdu, China	Urban road	GPS	Non-parametric	Vehicle trajectory data	LSTM Neural Network
2018	Gupta et al.	Porto, Portugal	Urban road	GPS	Non-parametric	Taxi travel speed	RF and gradient boosting
2019	Moonam et al.	Madison, Wisconsin, US	Freeway	Bluetooth detector	Non-parametric	Travel speed	KNN, KF
2019	Kumar et al.	Chennai, India	Urban road	GPS	Non-parametric	Travel time	KNN
2019	Cristobal et al.	Gran Canaria, Spain	Urban road	Public transport networkl	Non-parametric	Travel time	K-Medoid Clustering Technique
2020	Kwak & Geroliminis	California, US	Freeway	PeMS	Parametric	Travel time	Dynamic linear model
2020	Fu et al.	Beijing, Suzhou, Shenyang,China	Urban road	Ride-hailing platform	Non-parametric	Travel time	Graph attention network
2021	Chiabaut & Faitout	Lyon, French	Highway	Loop detector	Non-parametric	Travel time	PCA and Clustering

2.4. Summary

In summary, with the wide applications of big data in the field of transportation, different machine learning approaches have been deployed in the travel time prediction area. The methodologies include, but are not limited to, the following: SVM 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. RF and gradient boosting), etc. Table 2.1 provides a summary of the studies reviewed in chronological order. A comprehensive review and synthesis of the current and historical research related to ravel time prediction and machine learning-based travel time prediction methodologies have been discussed and presented in the preceding sections. This is intended to provide a solid reference and assistance in analyzing travel time and developing travel time prediction models for future tasks.

CHAPTER 3: DATA DESCRIPTION AND PROCESSING

3.1. Introduction

This chapter provides the basic information needed for 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

In this study, the travel time dataset is gathered from the Regional Integrated Transportation Information System (RITIS). RITIS is an advanced traffic system which includes the segment analysis, probe data analytics, and signal analytics. I-485 is one of the most heavily traveled interstate freeways in the City of Charlotte which loops encircling the city. A series of segments in the southern loop are selected for the case study. In order to achieve an acceptable accuracy of prediction, the model has to be well-established with large historical data that need to be secured and typically contain at least one-year's data (Torday, 2010). In this study, the dataset is collected from 01/01/2019-12/31/2019 and the time interval is 15 minutes, which have uninterrupted coverage in the RITIS data with 24 hours per day and 365 days a year. The selected study section starts from the interchange with I-77 (Exit 67) and ends at the interchange with US-74 (Exit 51). Figure 3.1 shows the study road segments and 3 traffic message record sensors (A, B, C) that were selected for the model validation. There are 37 miles of roadways and 32 traffic message channel code segments in the clockwise and counterclockwise directions.

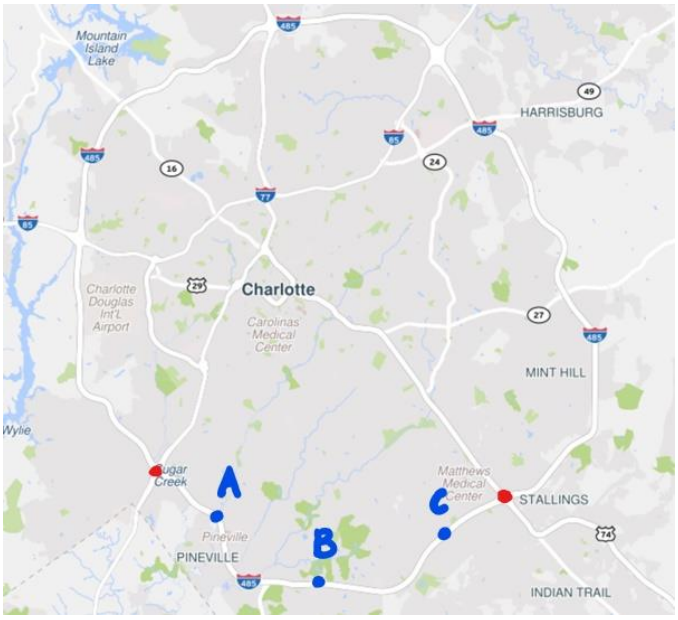

Figure 3.1 Selected Road Segment on Southern 485

In this study, the raw weather data are collected at locations that are close to the Charlotte Douglas International airport, which is not far from the selected roadway segments. The raw weather data include weather information such as temperature, dew point, humidity, pressure, visibility, wind direction, wind speed, gust speed, precipitation, and conditions. Table 3.1 shows a sample of raw weather data.

Table 3.1 Sample Raw Weather Data

Date	Time (EDT)	Conditions
Saturday, Oct 5th, 2019	7:55 AM	Rain
Saturday, Oct 5th, 2019	8:55 AM	Rain
Saturday, Oct 5th, 2019	9:55 AM	Light Rain
Saturday, Oct 5th, 2019	10:55 AM	Light Rain
Saturday, Oct 5th, 2019	11:55 AM	Light Rain
Saturday, Oct 5th, 2019	12:55 AM	Light Rain
Saturday, Oct 5th, 2019	13:55 PM	Light Rain

It was found that the travel time reliability is sensitive to the weather condition and severe weather (Zhao & Chien, 2012). The weather can greatly affect the travel time and speed, which are two important traffic flow parameters of transportation, resulting in deterioration of a traffic system's performance (Koetse & Rietveld, 2007). Since the weather data were recorded on a per hour basis, the discrepancy in the time intervals was treated by developing and using a mapping methodology to combine the traffic data with the weather data. The weather conditions were originally classified into 30 detailed weather conditions. In order to improve the computing power of the model, the weather conditions are further categorized into three groups (normal, rain, and snow/fog/ice) in this study. Table 3.2 presents the detailed classification of the newly grouped weather conditions.

Table 3.2 Classification of the Weather Conditions

Original Weather Condition	Weather Category in study
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	
Squalls	
Light Rain	Rain
Rain	
Heavy Rain	
Light Drizzle	
Heavy Thunderstorm	
Thunderstorms an	
Light Thunderstorm	
Thunderstorm	
Drizzle	

3.3. Feature Selection and Pre-Processing Steps

In this study, the dataset is collected from the southern part of the I-485 freeway, which is divided into 32 sections by the recorded sensor segment. The traffic information on each segment (from sensor to sensor) contains the subject segment and adjacent segment travel times,

Day of Time (DOW), Time of Day (TOD), segment length, and space mean speed and other traffic information. The missing data rate of the sample dataset in this study is less than 0.5% (i.e., 4246 out of 981,083), and the missing values are replaced with the mean of its closest surrounding values. A well-defined preprocessing capability that corrects various types of data errors including missing data is mandatory for a reliable travel time prediction system with acceptable accuracy and efficiency (Oh et al.,2018). In this study, prediction model is developed under normal traffic conditions and does not consider the factor of unexpected conditions (e.g., serious accidents or special events). Table 3.3 presents the definitions and attributes on selected features. Figure 3.2 shows the data pre-preparation process.

Table 3.3 Definitions and Attributes on Selected Features

Variable	Attribute	Definition
ID	Categorical	Road segment ID
T_t	Float	The travel time at the prediction road segment
Speed	Float	Space Mean Speed
TOD	Categorical	Time of day is indexed from 1 to 96, which represent the time from 0:00-24:00 by every 15-minute timestep
DOW	Categorical	Day of week is indexed from 1 to 7, which represent from Monday through Sunday
Month	Categorical	The Month is indexed 1 to 12, which represent from January to December
Weather	Categorical	Weather is indexed from 1 to 3, which represent normal, rain and snow/ice/fog
T_{t-1}	Float	The travel time at prediction segment 15 minutes before
T_{t-2}	Float	The travel time at prediction segment 30 minutes before
T_{t-w}	Float	The travel time at prediction segment 1 week before
ΔT_{t-1}	Float	The ravel time change value at T_{t-1}
ΔT_{t-2}	Float	The ravel time change value at T_{t-2}
ΔT_{t-w}	Float	The travel time change value at T_{t-w}
T_{t-1}^{i-1}	Float	The travel time of the nearest upstream road segment 15 minutes before
T_{t-1}^{i-2}	Float	The travel time of the second nearest upstream road segment 15 minutes before

ΔT_{t-1}^{i-1}	Float	The travel time change value at the nearest upstream road segment 15 minutes before
ΔT_{t-1}^{i-2}	Float	The travel time change value at the second nearest upstream road segment 15 minutes before
T_{t-1}^{i+1}	Float	The travel time of the nearest downstream road segment 15 minutes before
T_{t-1}^{i+2}	Float	The travel time of the second nearest downstream road segment 15 minutes before
ΔT_{t-1}^{i+1}	Float	The travel time change value at the nearest downstream road segment 15 minutes before
ΔT_{t-1}^{i+2}	Float	The travel time change value at the second nearest downstream road segment 15 minutes before

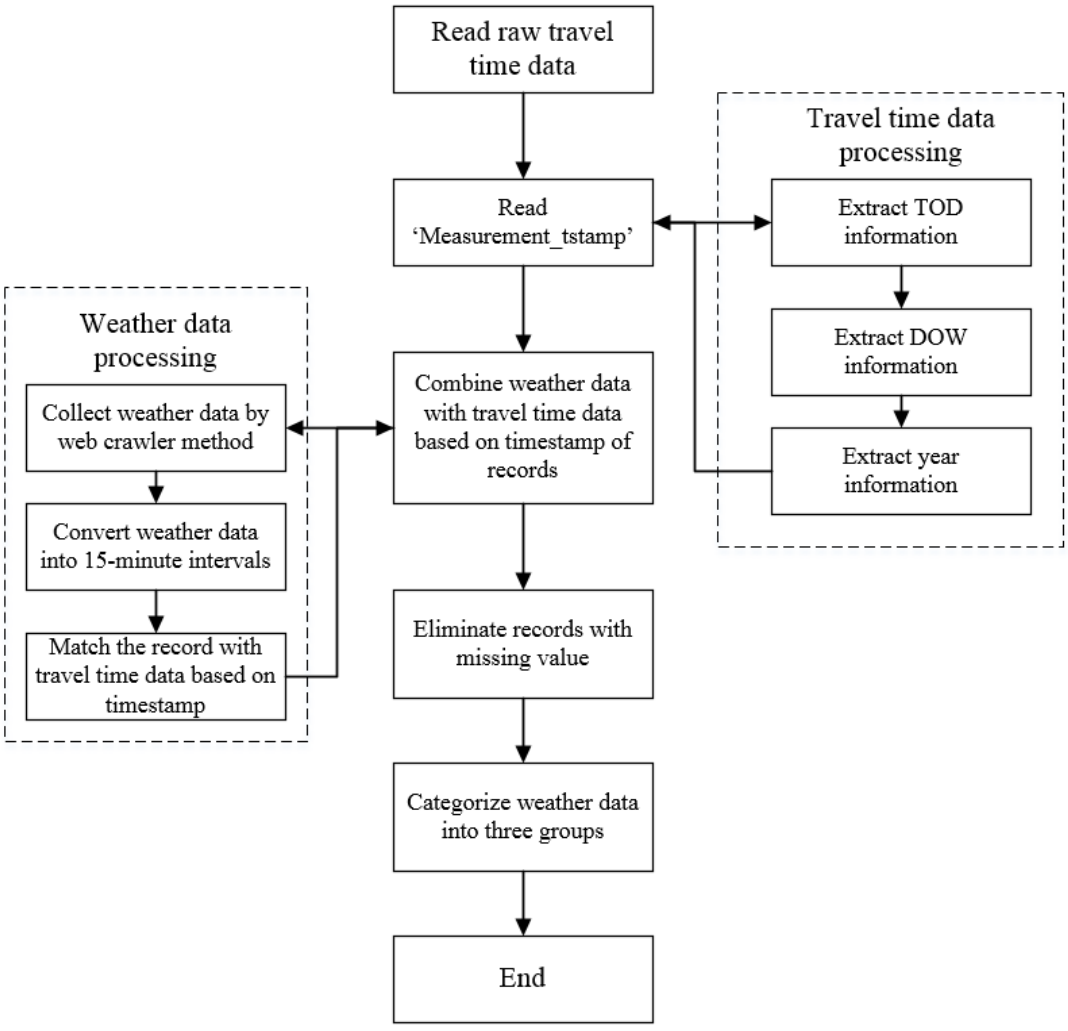

Figure 3.2 The Data Pre-preparation Steps

3.4. Summary

This chapter presents the detailed information on the data source collection, data structure, and pre-preparation approach to combine the travel time with raw weather data. This is intended to provide a solid reference and assistance in travel time prediction for future tasks.

CHAPTER 4: TRAVEL TIME PREDICTION METHODOLOGY

Chapter 4 will discuss the travel time prediction methodology based on the data described in Chapter 3. Two machine learning based travel time prediction models (RF and XGBoost) will be developed. The detailed modelling process of RF will be described such as the data structure configuration, parameter determination, model training, and model validation.

4.1. Random Forest Algorithm

4.1.1. Ensemble Learning Methodology

An ensemble itself is a supervised learning algorithm, which can be trained and used to make predictions. The ensemble learning-based algorithms consist of multiple base models (e.g. decision tree model), each of which provides an alternative solution to the problem. The prediction results tend to be more accurate when there is a strong diversity among the models (Kuncheva and Whitaker, 2003). Decision trees always suffer from high variance which causes the instability of the prediction results. Bootstrap aggregating (bagging) is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms. In the bagging process, the algorithm builds multiple models from the same original samples data set to reduce the variance. However, the bagging can make the trees highly correlated. RF is an extension of bagging in that in addition to building trees based on multiple samples of the original training data, it also constrains the features that can be used to build the trees, forcing trees to be different. To date, the RF models have been widely applied to various research fields (Greenhalgh and Mirmehdi, 2012; Xu et al., 2016). For classification tasks, RF typically gives high accuracy while also having a faster classification time. An RF classifier requires training with large data sets, which in our study are obviously available because of the nature of the travel record data collected. Furthermore, the RF computational process runs

efficiently on large data sets, which can reduce model complexity, overcome the overfitting to some extent and improve the efficiency. As known, overfitting means that the estimated model fits the training data too well. Generally, this is caused by the fact that the model function is too complicated to consider each data point and even outliers. The RF method can build a large number of random trees and then combine the results from each individual tree. The benefit of using the RF methods is that through averaging, the variance can be reduced.

4.1.2. Random Forest Algorithm

RF is an algorithm that can compete with gradient enhancement tree in integrated learning, especially for its convenient parallel training, which is very tempting in the era of big data and large samples. For each tree, the feature selection is conducted randomly. The prediction process is showed in the Figure 4.1. The difference between RF algorithm and the decision tree algorithm is that in RF, the processes of finding the root node and splitting the feature nodes will run randomly.

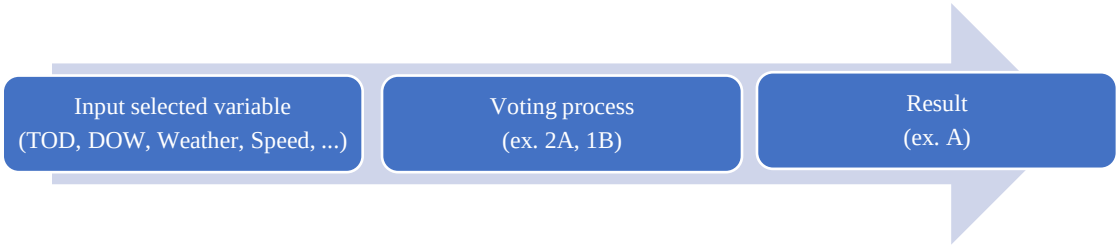

Figure 4.1 Prediction Process of RF Algorithm

Figure 4.2 shows the prediction process of RF algorithm, which is described as follows.

(1) The number of training data points is *N*, and the number of variables in the classifier is *M*.

(2) Select the *m* variables in the whole variable set *M* to determine the decision at a node of the tree. (Note that *m* is always considerably smaller than *M*)

(3) To construct the forest by trees, choose a training set *k* times with replacement from all *N* training dataset. Each of these datasets is called a bootstrap dataset. The number *k* is the number of the trees to be trained.

(4) For each tree node, randomly choose *m* variables on which to make the decision at that node. Calculate and get the best split based on these *m* variables in the training set.

(5) The "*Gini Index*" is used for calculating the Gini value to determine the best split point, which can be used to describe the purity after split. The *Gini index* will fall between 0 to 1 and the smaller the value, the better the split. If a dataset contains elements from two classes, the *Gini index* is defined as follows:

$$Gini(T) = 1 - \sum_{j=1}^{n} (p_j^2) \tag{1}$$

where p^j is the relative proportion of class *j* in the original dataset *T*, and *n* is the number of classes in dataset *T*.

$$Gini_{split}(T) = \frac{N_1}{N} Gini(T_1) + \frac{N_2}{N} Gini(T_2) \tag{2}$$

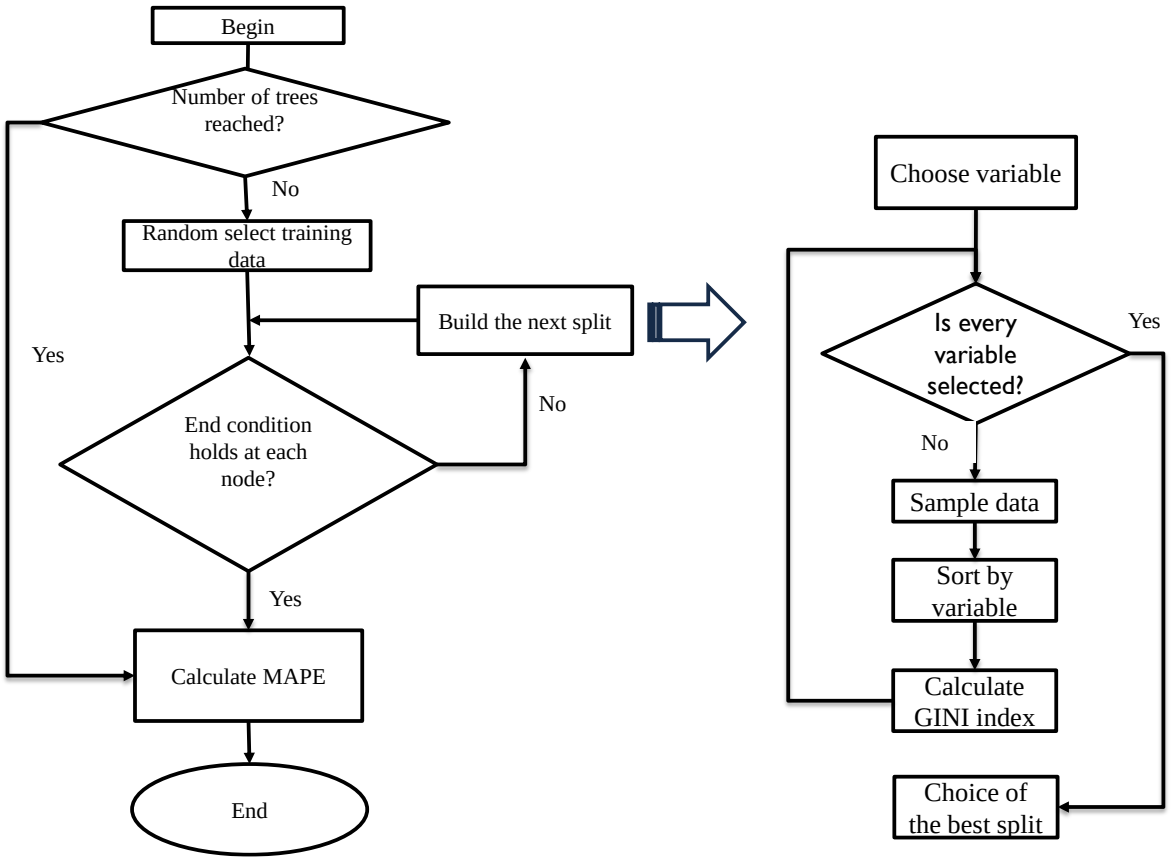

Figure 4.2 RF Algorithm Processing Flow

4.1.3. Proposed Travel Time Prediction Approaches

4.1.3.1. Feature Selection and Pre-Processing Steps

In the prediction model, the southern part of the I-485 freeway is divided into 32 sections by the recorded sensor segment in this study. Traffic data on each segment (from sensor to sensor) contains information on the subject segment and adjacent segment travel times, Day of Time (DOW), Time of Day (TOD), segment length, and space mean speed. The RITIS real-world travel time data used for this study have a less than 0.5% missing rate (i.e., 4246 out of 981,083). Note that in this study, the missing values are simply replaced with the mean of its closest surrounding values. From the previous studies (e.g., Wang et al., 2018), the variables that have a significant impact on the travel time prediction included the basic variables (such as time of day,

day of week, month and weather) and the spatial and temporal characteristics of the adjacent road segments. Furthermore, in this study, the travel times (which are collected several steps ahead of the travel time to be predicted) are also accounted for in the model estimation. The prediction model is developed under normal traffic conditions and does not consider unexpected conditions (e.g., special events). The data on each segment will be used to train one forest which consists of decision trees. The RF model prediction includes two major steps: training and prediction. The forests are constructed by using randomly selected parameter combinations and different numbers of trees during the training step. Table 4.1 presents the definitions and attributes on selected features that are used in this study.

Table 4.1 Definitions and Attributes on Selected Features

Variable	Definition	Attribute
ID	Road segment ID	Categorical
L	Length of the road segment	Float
Speed	Space Mean Speed	Float
TOD	Time of day is indexed from 1 to 96, which represent the time from 0:00-24:00 by every 15-minute timestep	Categorical
DOW	Day of week is indexed from 1 to 7, which represent from Monday through Sunday	Categorical
Month	The Month is indexed 1 to 12, which represent from January to December	Categorical
Weather	Weather is indexed from 1 to 3, which represent normal, rain and snow/ice/fog	Categorical
T_{t-1}	The travel time at prediction segment 15 minutes before	Float
T_{t-2}	The travel time at prediction segment 30 minutes before	Float
T_{t-3}	The travel time at prediction segment 45 minutes before	Float
T_{t-w}	The travel time at prediction segment 1 week before	Float
ΔT_{t-1}	The ravel time change value at T_{t-1}	Float
ΔT_{t-2}	The ravel time change value at T_{t-2}	Float
ΔT_{t-3}	The travel time change value at T_{t-3}	Float
ΔT_{t-w}	The travel time change value at T_{t-w}	Float
T_{t-1}^{i-1}	The travel time of the nearest upstream road segment 15 minutes before	Float

T_{t-1}^{i-2}	The travel time of the second nearest upstream road segment 15 minutes before	Float
ΔT_{t-1}^{i-1}	The travel time change value at the nearest upstream road segment 15 minutes before	Float
ΔT_{t-1}^{i-2}	The travel time change value at the second nearest upstream road segment 15 minutes before	Float
T_{t-1}^{i+1}	The travel time of the nearest downstream road segment 15 minutes before	Float
T_{t-1}^{i+2}	The travel time of the second nearest downstream road segment 15 minutes before	Float
ΔT_{t-1}^{i+1}	The travel time change value at the nearest downstream road segment 15 minutes before	Float
ΔT_{t-1}^{i+2}	The travel time change value at the second nearest downstream road segment 15 minutes before	Float
T_t	The travel time at the prediction road segment	Float

4.1.3.2. The Coefficients in the RF Model and the Parameters' Tuning Process

To achieve the best modelling results, it is important to explore the effect of different combinations of parameters on the RF model prediction performance. Based on previous studies, there are primarily three features that can be tuned to optimize the predictive power of the model: *Max_features, N_estimators* (number of trees), and *Min_sample-leaf*. They are presented as follows:

***Max_features*:**

This is the maximum number of features in the RF model that is allowed to try in each individual tree. There are multiple options available in Python to assign maximum features. "Auto/None" is a command that simply takes all the features that make sense in every tree, which simply does not put any restrictions on the individual tree. The "SQRT" option takes square root of the total number of features in each individual run. For example, if the total number of variables is 100, under this option the system can only take 10 of them in each individual tree. The "log2" option is another similar type of option used for *max_features*. In this study, after several tests, the random subspace method is applied. The number of features considered at each internal node of

random forests is *m*, which is randomly chosen to be $m = INT(log_2 M + 1)$, where *m* is the total number of features, as suggested by Breiman (2001a, b).

***n_estimators*:**

This is the number of trees that one wants to build before taking the maximum voting or averages of predictions. A larger number of trees will give one better performance with a compromise of computing efficiency. As such, one should choose a value as high as what the processor can handle because this makes the predictions stronger and more stable.

***min_sample_leaf*:**

Leaf is the end node of a decision tree. A smaller leaf makes the model more prone to capture noise in the train data. In this study, after several trials of different leaf size, a minimum leaf size of 20 is chosen. In addition, researchers have to face the problem named "tuning RF parameters in practice", the good answer to which varies from dataset to dataset. In this study, the tool RandomSearch is applied to optimize the tuning process. To do so, one needs to define the range of parameters and then run these procedures to get the best model. In this study, the first run is 1000 trees, with 1/2 features per node. RF models are not sensitive if the features are independent or dependent, though many will perform better if the data are preprocessed. A simple way to identify dependence among features is to calculate a correlation coefficient between each feature and all other features. To identify the importance of the features, one can build a forest and see which features get used, as RF models tend to split out the results by using the most statistically significant features.

It is also important to note that the performance measure used in this study is the MAPE. The statistic mean absolute percentage error usually expresses accuracy as a percentage that is calculated as follows:

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

where, m = The total number of the data points,

$\hat{y_i}$ = The predicted travel time value in the test dataset of record *i*,

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

4.2. Summary

This chaper describes the RF travel time prediction methodology, which includes the discussions of the RF algorithm, the parameter determination and model training process. To optimize the predictive power of the model, there are three features that need to be tuned: *Max_features, N_estimators* (number of trees), and *Min_sample-leaf*. The potential features from the dataset include but are not limited to: time of day, day of week, month of year, weather conditions, segment characteristics, etc. The RF based travel time prediction model is developed based on the features that are generated from the original data ascdescribed in previous chapter. The detail of the coefficients in the RF model and the parameters tuning process is described. The statistic MAPE is discribed and selected as the prediction model performance measurement, which means the MAPE will be used to measure the prediction error.

CHAPTER 5: TRAVEL TIME PREDICTION MODEL VALIDATION

5.1. Introduction

Chapter 5 will present the validation of the proposed machine learning models based on the data described in Chapter 4. For the RF model, the data training step will be described to determine the parameters in the model structure. The potential features will include but not limit to: time of day, day of week, month of year, weather conditions, segment characteristics, etc. One of the most widely used criteria MAPE will be used to measure the prediction error.

5.2. The Tunning and Validation Process of RF Model

As mentioned in chapter 4, to improve and finally achive the best modelling results, we need to explore the effect of different combinations of parameters on the RF model. Based on previous studies, three features (Max_features, N_estimators (number of trees), and Min_sample-leaf) need to be tuned to optimize the prediction result.

From the Figure 5.1 and Table 5.1, the tunning process shows that when the number of trees reaches 50, the value of MAPE becomes nearly the same. In statistics, overfitting is the co-product of an analysis that corresponds too exactly to the sample set of data, and therefore, may fail to fit additional data or predict future observations reliably, which 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 tree base model (Zhang and Haghani, 2015). There is also a need to consider the tradeoff between prediction accuracy and computational time. Since a large number of trees are being fitted, model complexity also increases and requires more computational time.

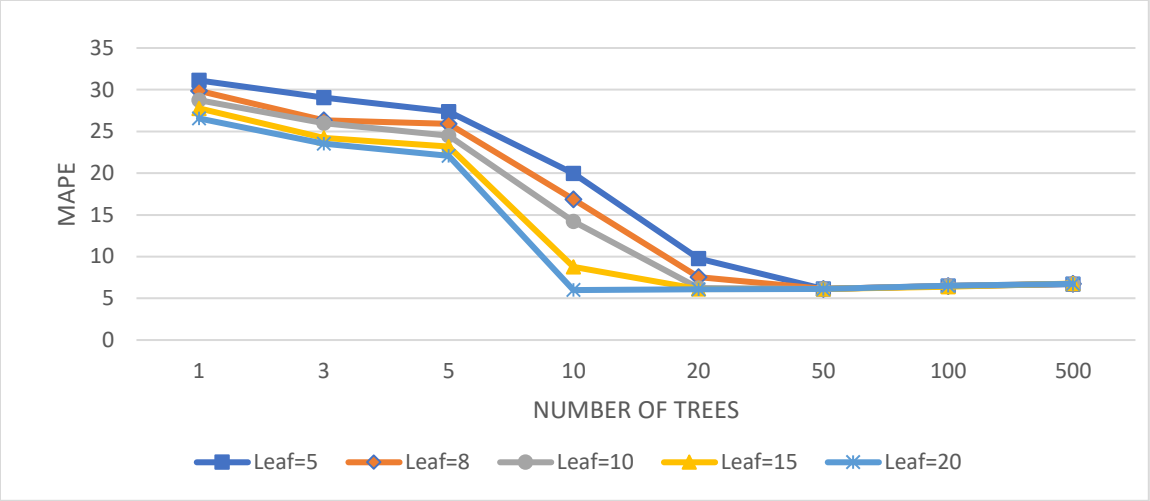

Figure 5.1 RF Travel Time Prediction Model Performance

Table 5.1 The MAPE of a Combination of Parameters

	MAPE				
Number of trees	Leaf=5	Leaf=8	Leaf=10	Leaf=15	Leaf=20
1	31.11	29.87	28.75	27.77	26.56
3	29.05	26.34	25.98	24.21	23.52
5	27.38	25.9	24.48	23.19	22.09
10	19.98	16.87	14.22	8.76	5.99
20	9.78	7.56	6.23	6.11	6.1
50	6.13	6.14	6.12	6.12	6.12
100	6.46	6.48	6.49	6.41	6.5
500	6.7	6.72	6.75	6.76	6.73

In the machine learning area, the predictor variables usually have significant impacts on the prediction results. Exploring the influence on the individual feature can help understand the variables better. Higher relative importance indicates a higher influence in predicting travel time. Table 5.2 presents the relative importance of each variable and their ranks in the optimized RF model. In the Table 5.2, each predictor variable has a different impact on the predicted travel time. The model result shows that the variable T_{t-1}(travel time 15 minutes before) contributes the most (34.85%) to the predicted travel time result. 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. TOD is the second highest ranked variable with the relative importance value of 30.12%, and this result is also under expectation. T_{t-w} is the fourth highest ranked variable with the importance value of 9.87%, which can be interpreted as a highly similar pattern of traffic times between weeks.

The result in Table 5.2 also shows that the spatial impact is less than the time impact since the relative importance values of all the spatial variables are less than 1% (except the variable road ID with a relative importance value of 2.28%). Several variables such as the travel time of the two upstream segments (with the relative importance value of 0.31% and 0.42%, respectively) and the travel time of the two downstream segments (with the relative importance value of 0.35% and 0.61%, 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 0.29%, and the relative importance values of the two upstream segments are 0.79% and 0.37%, respectively. Based on these results, it could be explained that the relative importance values of the downstream segments are higher than those of upstream segments. The reason is caused by the spatial characteristics of the roadway. When a bottleneck occurs at the downstream segment, the upstream will be impacted very shortly.

Table 5.2 Relative Importance of Each Variable and their Ranks in the RF Model

Variable	Definition	Relative Importance (%)	Attribute
ID	Road segment ID	2.28	7
L	Length of the road segment	0.17	23
Speed	Space Mean Speed	10.59	3
TOD	Time of day is indexed from 1 to 96, which represent the time from 0:00-24:00 by every 15-minute timestep	30.12	2
DOW	Day of week is indexed from 1 to 7, which represent from Monday through Sunday	2.84	5
Month	The Month is indexed 1 to 12, which represent from January to December	1.59	8
Weather	Weather is indexed from 1 to 3, which represent normal, rain and snow/ice/fog	2.63	6
T_{t-1}	The travel time at prediction segment 15 minutes before	34.85	1
T_{t-2}	The travel time at prediction segment 30 minutes before	0.57	11
T_{t-3}	The travel time at prediction segment 45 minutes before	0.28	18
T_{t-w}	The travel time at prediction segment 1 week before	9.87	4
ΔT_{t-1}	The ravel time change value at T_{t-1}	0.24	19
ΔT_{t-2}	The ravel time change value at T_{t-2}	0.20	21
ΔT_{t-3}	The travel time change value at T_{t-3}	0.18	22
ΔT_{t-w}	The travel time change value at T_{t-w}	0.22	20
T_{t-1}^{i-1}	The travel time of the nearest upstream road segment 15 minutes before	0.31	15
T_{t-1}^{i-2}	The travel time of the second nearest upstream road segment 15 minutes before	0.42	12
ΔT_{t-1}^{i-1}	The travel time change value at the nearest upstream road segment 15 minutes before	0.29	16
ΔT_{t-1}^{i-2}	The travel time change value at the second nearest upstream road segment 15 minutes before	0.29	16
T_{t-1}^{i+1}	The travel time of the nearest downstream road segment 15 minutes before	0.35	14
T_{t-1}^{i+2}	The travel time of the second nearest downstream road segment 15 minutes before	0.61	10
ΔT_{t-1}^{i+1}	The travel time change value at the nearest downstream road segment 15 minutes before	0.79	9
ΔT_{t-1}^{i+2}	The travel time change value at the second nearest downstream road segment 15 minutes before	0.37	13

To measure the effectiveness of different travel time prediction algorithms, the MAPEs are computed for 3 different observation segments (A, B, C are three observation segments along the selected study freeway, shows in Figure 5.2) with different prediction horizon from 15 minutes to 60 minutes. According to the comparison shown in Table 5.3 and Figure 5.3, the performance of the proposed RF is better than the eXtreme Gradient Boosting (XGBoost, another widely used tree-based ensemble method), especially when the horizon of prediction time is long. The MAPEs of RF model are significantly smaller than XGBoost when the horizon is long enough (i.e., longer than 45 min).

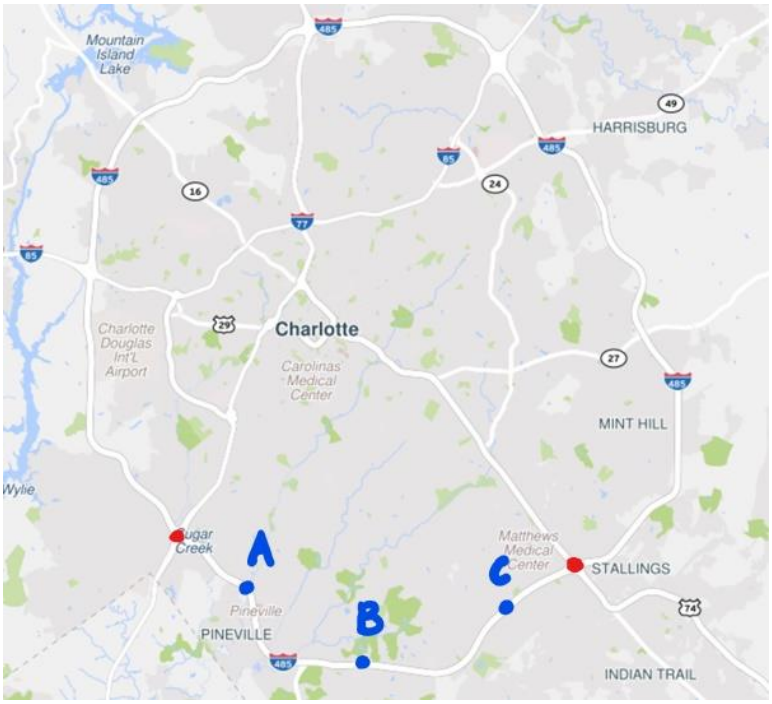

Figure 5.2 Study Road Segment

Table 5.3 The Comparison of Different Prediction Method

MAPE (%) of different observation point with different prediction time range												
Models	15min			30min			45min			60min		
	A	B	C	A	B	C	A	B	C	A	B	C
RF	6.49	6.15	6.39	9.69	9.97	10.67	15.29	16.19	17.37	24.59	25.66	26.76
XGBoost	6.57	6.14	6.39	10.58	9.98	10.89	15.35	15.98	17.90	25.90	26.06	28.09

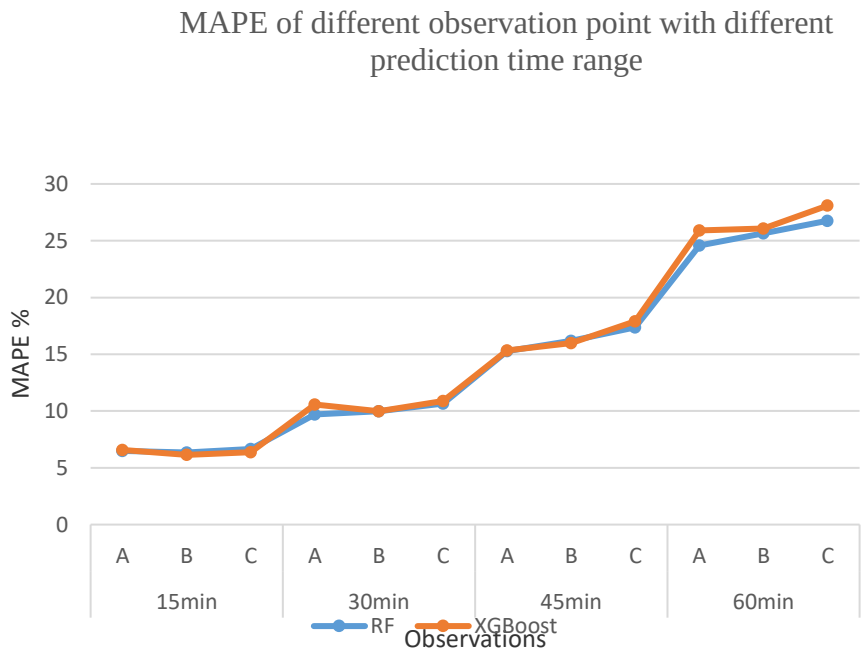

Figure 5.3 MAPE of Different Observation Point with Different Prediction Time Range

5.3. Summary

In summary, we presents the validation of the proposed machine learning models based on the sample dataset. For the machine learning prediction model, the data training step will be described to determine the parameters in the model structure. The relative importance of each variable (selected feature) were ranked in the RF travel time prediction model. The relative importance of the features shows that the travel time one step ahead (15 minutes before) contributes the most (34.85%) to the predicted travel time. The time of day, day of the week, the travel time along the prediction segment one week before, and weather also have higher relative

importance values in the model than other features. Adding up the most important six variables' relative importance values (T_{t-1}, TOD, Speed, T_{t-w}, DOW, Weather) in Table 5.2 is as high as 90.90%, which means that these six selected variables include most of the information needed in the travel time prediction. Table 5.2 also shows that the time features (such as T_{t-1}, TOD, T_{t-w}, DOW) have significantly higher relative importance values than the spatial and weather features (such as weather, road ID, length, and speed). According to the the prediction performance comparison from different selected road test points and different prediction time horizons, the proposed RF is proved better than the XGBoost especially when the horizon of prediction time is long. The MAPEs of RF model are significantly smaller than XGBoost when the horizon is longer than 45 min.

CHAPTER 6: SUMMARY AND FUTURE RESEARCH

6.1. Summary

Travel time prediction is based on accurate modeling the complex non-linear spatiotemporal traffic dynamics in the real world (Ran et al., 2019). The accuracy and interpretability of models are two major concerns. In general, RF more like a complex black box models for accuracy versus less accurate but more interpretable traditional models such as linear regression. In recent year, the increase congestion on freeways has led the increasing of uncertainty, which made the TTP model more difficult to achieve the preset prediction accuracy. In this paper, a systematic machine learning solution is proposed for short-term TTP. Short-term travel time prediction can be an important planning tool for both individuals and public transportation. In both cases, it is expected that the application of accurate travel time prediction can improve improve the level of service and travel planning by reducing errors between the actual and predicted travel time, which will also reduce the whole cost of travel and deliveries. The tree-based ensemble methods have been widely used in the field of prediction. By combining a simple tree to a forest, RF always produces high prediction accuracy (Zhang and Haghani, 2015). In this research, RF approach was developed and compared with XGBoost. The impact of the pre-processing tasks (feature selection, domain values definition) in different methods on their performance was also studied. Sample dataset from I-485 charlotte was selected to conduct a case study and experiments indicated that RF is the most promising approach among all algorithms tested. The results showed that all ensemble learning methods achieve a high estimation accuracy.

Most existing machine learning models can capture the nonlinear pattern of travel time but suffer from over-fitting. Study results indicate that the RF model has its considerable advantages in freeway travel time prediction, and the performance evaluation result also shows that the RF-

based model can have better predictions in terms of prediction accuracy. RF model showed a reasonable performance compared with other approaches. When the prediction time is no more than 15 mins, the RF algorithm is relatively accurate. However, when the prediction horizon is longer than 30 minutes, the error increases dramatically. Different from other machine learning methods, RF methods provide interpretable results with different types of predictor variables. RF can also handle data with very high dimensions (many features) without feature selection (because feature subsets are randomly selected), and identifies which features are more important after training process. Furthermore, it has an effective way of estimating missing data and maintaining accuracy when a significant proportion of the data is missing. 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 the travel time at prediction segment one week before and weather also have higher relative importance values in the model than other features. Adding up the most important eight variables' relative importance values (T_{t-1}, TOD, Speed, $T_{t-\mathrm{w}}$, DOW, Weather, Road ID, Month) will be as high as 94.77%, which means that these eight selected variables include most of the information needed in the travel time prediction. The proposed RF travel time prediction method has considerable advantages over the other tree-based approach. However, there are still some limitations associated with this RF approach. For example, the prediction model is developed under normal traffic conditions and does not consider unexpected conditions (e.g., special events).

6.2. Future Research

As mentioned, the practice of RF algorithm and other tree-based ensemble methods in travel time prediction area are still very limited. The future focus of the research would be hybrid models (combination model), which can combine several models of the same or different types of prediction models to enhance the model performance and prediction. The RF method can be combined with other tree-based methods or another type of machine learning method in the preprocessing step or prediction step. Experimental results showed the combination methods have a better prediction result then using a method alone (Li et al., 2009). As the combination model method has been proved superior in terms of prediction accuracy, this should be given careful consideration in the future.

REFERENCES

1. Ahmed, M.S. and Cook, A.R., 1979. Analysis of freeway traffic time-series data by using Box-Jenkins techniques. Transportation Research Record, 722, 1-9.
2. Bengio, Y., Simard, P. and Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks, 5(2), pp.157-166.
3. Breiman, L., 2001. Random forests, machine learning 45. J. Clin. Microbiol, 2(30), 199-228.
4. Cristobal, T., Padron, G., Quesada-Arencibia, A., Alayon, F., de Blasio, G. and García, C.R. (2019), "Bus travel time prediction model based on profile similarity", Sensors, Vol. 19 No. 13, p. 2869.
5. Dharia, A., & Adeli, H., 2003. Neural network model for rapid forecasting of freeway link travel time. Engineering Applications of Artificial Intelligence, 16(7-8), 607-613.
6. Dion, F., Rakha, H., & Kang, Y. S., 2004. Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections. Transportation Research Part B: Methodological, 38(2), 99-122.
7. Du, L., Peeta, S., & Kim, Y. H., 2012. An adaptive information fusion model to predict the short-term link travel time distribution in dynamic traffic networks. Transportation Research Part B: Methodological, 46(1), 235-252.
8. Duan, Y., Yisheng, L.V. and Wang, F.Y., 2016, November. Travel time prediction with LSTM neural network. In 2016 IEEE 19th international conference on intelligent transportation systems (ITSC) (pp. 1053-1058). IEEE.
9. Fan, S. K. S., Su, C. J., Nien, H. T., Tsai, P. F., & Cheng, C. Y., 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.

10. Fu, K., Meng, F., Ye, J. and Wang, Z. (2020), “CompactETA: a fast inference system for travel time prediction”, Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 3337-3345.
11. Greenhalgh, J. and Mirmehdi, M., 2012. Real-time detection and recognition of road traffic signs. IEEE transactions on intelligent transportation systems, 13(4), pp.1498-1506.
12. Gupta, B., Awasthi, S., Gupta, R., Ram, L., Kumar, P., Prasad, B. R., & Agarwal, S., 2018. Taxi travel time prediction using ensemble-based random forest and gradient boosting model. In Advances in Big Data and Cloud Computing (pp. 63-78). Springer, Singapore.
13. Hamner, B., 2010. Predicting travel times with context-dependent random forests by modeling local and aggregate traffic flow. In 2010 IEEE International Conference on Data Mining Workshops (pp. 1357-1359). IEEE.
14. He, H. S., Keane, R. E., & Iverson, L. R., 2008. Forest landscape models, a tool for understanding the effect of the large-scale and long-term landscape processes. Forest Ecology and Management. 254: 371-374., 254.
15. Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.
16. Jiang, X., & Adeli, H., 2004. Wavelet packet‐autocorrelation function method for traffic flow pattern analysis. Computer‐Aided Civil and Infrastructure Engineering, 19(5), 324-337.
17. Koetse, M.J. and Rietveld, P., 2007, September. Climate change, adverse weather conditions, and transport: A literature survey. In Proceedings of the 9th NECTAR Conference. Network on European Communication and Transportation Activities Research (NECTAR), Porto, Portugal (CDROM).

18. Kuncheva, L. I., & Whitaker, C. J., 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine learning, 51(2), 181-207.
19. Kankanamge, K.D., Witharanage, Y.R., Withanage, C.S., Hansini, M., Lakmal, D. and Thayasivam, U., 2019, July. Taxi trip trael time prediction with isolated XGBoost regression. In 2019 Moratuwa Engineering Research Conference (MERCon) (pp. 54-59). IEEE.
20. Kwak, S. and Geroliminis, N. (2020), "Travel time prediction for congested freeways with a dynamic linear model", IEEE Transactions on Intelligent Transportation Systems.
21. Kumar, B.A., Jairam, R., Arkatkar, S.S. and Vanajakshi, L. (2019), "Real time bus travel time prediction using k-NN classifier", Transportation Letters, Vol. 11 No. 7, pp. 362-372.
22. Leshem, G., & Ritov, Y., 2007. Traffic flow prediction using adaboost algorithm with random forests as a weak learner. In Proceedings of world academy of science, engineering and technology (Vol. 19, pp. 193-198). Citeseer.
23. Li, Y., Fujimoto, R. M., & Hunter, M. P., 2009. Online travel time prediction based on boosting. In 2009 12th International IEEE Conference on Intelligent Transportation Systems (pp. 1-6). IEEE.
24. Liu, Y., Wang, Y., Yang, X. and Zhang, L., 2017, October. Short-term travel time prediction by deep learning: a comparison of different LSTM-DNN models. In 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC) (pp. 1-8). IEEE.
25. Lee, E.H., Kho, S.Y., Kim, D.K. and Cho, S.H., 2020, July. Travel time prediction using gated recurrent unit and spatio-temporal algorithm. In Proceedings of the Institution of Civil Engineers-Municipal Engineer (pp. 1-9). Thomas Telford Ltd.

26. Ma, X., Tao, Z., Wang, Y., Yu, H. and Wang, Y., 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C: Emerging Technologies, 54, pp.187-197.
27. Mendes-Moreira, J., Jorge, A. M., de Sousa, J. F., & Soares, C., 2015. Improving the accuracy of long-term travel time prediction using heterogeneous ensembles. Neurocomputing, 150, 428-439.
28. Mori, U., Mendiburu, A., Álvarez, M. and Lozano, J.A., 2015. A review of travel time estimation and forecasting for advanced traveller information systems. Transportmetrica A: Transport Science, 11(2), pp.119-157.
29. Oh, S., Byon, Y. J., Jang, K., & Yeo, H., 2015. Short-term travel-time prediction on highway: a review of the data-driven approach. Transport Reviews, 35(1), 4-32.
30. Oh, S., Byon, Y.J., Jang, K. and Yeo, H., 2018. Short-term travel-time prediction on highway: A review on model-based approach. KSCE Journal of Civil Engineering, 22(1), pp.298-310.
31. Papageorgiou, M., Papamichail, I., Messmer, A., & Wang, Y., 2010. Traffic simulation with METANET. In Fundamentals of traffic simulation (pp. 399-430). Springer, New York, NY.
32. Park, D. and Rilett, L.R., 1999. Forecasting freeway link travel times with a multilayer feedforward neural network. Computer-Aided Civil and Infrastructure Engineering, 14(5), pp.357-367.
33. Ramezani, M., & Geroliminis, N., 2012. On the estimation of arterial route travel time distribution with Markov chains. Transportation Research Part B: Methodological, 46(10), 1576-1590.

34. Ran, X., Shan, Z., Fang, Y. and Lin, C., 2019. An LSTM-based method with attention mechanism for travel time prediction. Sensors, 19(4), p.861.
35. Ruppert, D., Wand, M.P. and Carroll, R.J., 2003. Semiparametric regression (No. 12). Cambridge university press.
36. Rupnik, J., Davies, J., Fortuna, B., Duke, A. and Clarke, S.S., 2015, October. Travel time prediction on highways. In *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing* (pp. 1435-1442). IEEE.
37. Ran, X., Shan, Z., Fang, Y. and Lin, C., 2019. An LSTM-based method with attention mechanism for TTP. Sensors, 19(4), p.861.
38. Schmitt, E. J., & Jula, H., 2007. On the limitations of linear models in predicting travel times. In 2007 IEEE Intelligent Transportation Systems Conference (pp. 830-835). IEEE.
39. Šemanjski, I., 2015. Analysed potential of big data and supervised machine learning techniques in effectively forecasting travel times from fused data. Promet-Traffic&Transportation, 27(6), 515-528.
40. Shen, L., 2008. Freeway travel time estimation and prediction using dynamic neural networks. Florida International University Electronic Theses and Dissertations, Miami, FL, U.S.A.
41. Sun, H., Liu, H. X., Xiao, H., & Ran, B., 2002. Short term traffic forecasting using the local linear regression model.

42. Torday, A., 2010. Simulation-based decision support system for real time traffic management (No. 10-2120). Transportation Research Board 89th annual meeting, Washington DC, United States.
43. Van Lint, J. W., 2006. Reliable real-time framework for short-term freeway travel time prediction. Journal of transportation engineering, 132(12), 921-932.
44. Van Lint, J.W.C., 2004. Reliable travel time prediction for freeways. Netherlands TRAIL Research School.
45. Vlahogianni, E. I., Golias, J. C., & Karlaftis, M. G., 2004. Short‐term traffic forecasting: Overview of objectives and methods. Transport reviews, 24(5), 533-557.
46. Wang, Z., Fu, K. and Ye, J., 2018, July. Learning to estimate the travel time. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 858-866).
47. Wu, C.H., Ho, J.M. and Lee, D.T., 2004. Travel-time prediction with support vector regression. IEEE transactions on intelligent transportation systems, 5(4), pp.276-281.
48. Wunderlich, K. E., Kaufman, D. E., & Smith, R. L., 2000. Link travel time prediction for decentralized route guidance architectures. IEEE Transactions on Intelligent Transportation Systems, 1(1), 4-14.
49. Wu, Z., Rilett, L.R. and Ren, W. (2021), "New methodologies for predicting corridor travel time mean and reliability", International Journal of Urban Sciences, pp. 1-24.
50. Xu, B. and Qiu, G., 2016, March. Crowd density estimation based on rich features and random projection forest. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1-8). IEEE.

51. Wang, Z., Fu, K. and Ye, J., 2018, July. Learning to estimate the travel time. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 858-866).
52. Yildirimoglu, M., & Geroliminis, N., 2013. Experienced travel time prediction for congested freeways. Transportation Research Part B: Methodological, 53, 45-63.
53. Yu, B., Wang, H., Shan, W., & Yao, B., 2018. Prediction of bus travel time using random forests based on near neighbors. Computer - Aided Civil and Infrastructure Engineering, 33(4), 333-350.
54. Zhang, Y., & Haghani, A.,2015. A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies, 58, 308-324.
55. Zhao, L. and Chien, S.I.J., 2012. Analysis of weather impact on travel speed and travel time reliability. In CICTP 2012: Multimodal Transportation Systems—Convenient, Safe, Cost-Effective, Efficient (pp. 1145-1155).
56. Zhu, T., Kong, X., & Lv, W., 2009. Large-scale travel time prediction for urban arterial roads based on Kalman filter. In 2009 International Conference on Computational Intelligence and Software Engineering (pp. 1-5). IEEE.
57. Zou, N., Wang, J., & Chang, G. L., 2008. A reliable hybrid prediction model for real-time travel time prediction with widely spaced detectors. In 2008 11th International IEEE Conference on Intelligent Transportation Systems (pp. 91-96). IEEE.
58. Zhao, Z., Chen, W., Wu, X., Chen, P.C. and Liu, J., 2017. LSTM network: a deep learning approach for short-term traffic forecast. IET Intelligent Transport Systems, 11(2), pp.68-75.